DOCUMENT RESUME

ED 193 955 FL 011 933

AUTHOR Horst, D. P.; And Others
TITLE An Evaluation of Project Information Packages (PIPs) as Used for the Diffusion of Bilingual Projects. Volume III: A Prototype Guide to Measuring Achievement Level and Program Impact on Achievement in Bilingual Projects.
INSTITUTION RMC Research Corp., Mountain View, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Office of Evaluation and Dissemination.
REPORT NO RMC-UR-460
PUB DATE May 80
CONTRACT 300-77-0313
NOTE 222p.; For related documents, see FL 011 931-932.

EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS *Academic Achievement; Bilingual Education; *Diffusion; *Program Effectiveness; Summative Evaluation
IDENTIFIERS *Bilingual Programs; *Project Information Packages

ABSTRACT

This report describes an evaluation of Project Information Packages (PIPs), sets of manuals and other materials intended to help a school district adopt and implement an exemplary education project. Four PIPs were evaluated in a field test, each PIP describing a different bilingual project. It was concluded that the awareness materials produced few application for PIPs. Field-test sites that received PIPs tended not to follow PIP guidelines closely, but to adapt them extensively, often with good justification. The bilingual programs at the sites were collectively successful, but the dissemination effort could not be judged a success. The present volume is a collection of specific evaluation guidelines and job aides that were developed for the use of the field-test sites and that have been organized in the format of a Prototype Evaluation Manual. This volume should be viewed as a preliminary draft. It deals in detail only with the evaluation of student achievement. (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED193955

# AN EVALUATION OF PROJECT INFORMATION PACKAGES (PIPs) AS USED FOR THE DIFFUSION OF BILINGUAL PROJECTS

## VOLUME III

## A PROTOTYPE GUIDE TO MEASURING ACHIEVEMENT LEVEL AND PROGRAM IMPACT ON ACHIEVEMENT IN BILINGUAL PROJECTS

D. P. Horst
D. M. Johnson
H. G. Nava
D. E. Douglas
L. D. Friendly
A. O. H. Roberts

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

RMC Report UR 460

USOE Contract No. 300-77-0313

May 1980

RMC Research Corporation
Mountain View, California

FL 011933

This report is made pursuant to contract No. 300-77-0313. The amount charged to the Department of Health, Education, and Welfare for the work resulting in this report (inclusive of the amounts so charged for any prior reports submitted under this contract) is $1,113,205. The names of the persons, employed or retained by the contractor, with managerial or professional responsibility for such work, or for the content of the report, are as follows:

D. P. Horst
D. E. Douglas
L. D. Friendly
D. M. Johnson
L. M. Luber
M. McKay
H. G. Nava
A. M. Piestrup
A. O. H. Roberts
A. Valdez

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

## PREFACE

This report describes an evaluation of Project Information Packages (PIPs), a specific type of packaging, as field tested by the United States Office of Education (USOE) for the diffusion of four bilingual projects. The field test began with the dissemination of the PIPs in the fall of 1976. The evaluation described here began about nine months later (summer, 1977) and continued through the 1978-1979 school year.

This report consists of three volumes, as follows:

Volume I, the Summary Report, comprises (a) an executive summary of the study questions and findings, (b) an introduction to the study (Section 1), (c) a non-technical summary of Substudy I, the process evaluation of the PIP diffusion effort (Section 2), and (d) a non-technical summary of Substudy II, the evaluation of the impact of the diffusion effort on students (Section 3). This volume is intended to provide a self-contained overview of the policy-related study questions and conclusions.

Volume II, the Technical Discussion and Appendices, documents the methodology and results of the two substudies and provides more detailed discussions of conclusions and recommendations. This volume also includes five appendices: (a) site-by-site results of the process substudy, (b) site-by-site results of the impact substudy, (c) the complete conceptual framework used in the process evaluation substudy, (d) a comparative analysis of the contents of the four bilingual PIPs, and (e) a summary of the major, mid-study inputs from the study advisory panel.

Volume III, the present volume, is a collection of specific evaluation guidelines and job aides that were developed for the use of the field-test sites and which have been organized in the format of a Prototype Evaluation Manual. This volume should be viewed as a preliminary draft rather than a finished product. Further, it deals in detail only with the evaluation of student achievement, which is only one component of a complete, bilingual program evaluation.

## NOTE

This volume comprises many of the specific achievement-evaluation guides and worksheets developed for the field-test sites in the bilingual-PIP evaluation. These materials are presented here in the form of an evaluation manual, in order to highlight the major issues that arose in the course of the field test and to illustrate the kinds of solutions that RMC has proposed.

This prototype manual is intended primarily to stimulate discussion on ways to resolve the major, practical problems in bilingual program evaluation. Many technical issues remain to be settled and, in addition, an extensive process of format development, tryout, and revision would be required before this manual could be considered ready for general use. Nevertheless, until more comprehensive, specific, evaluation guidelines for bilingual programs are developed, we believe that these materials may be of use to some bilingual-program personnel.

CONTENTS

AN ORIENTATION TO THE EVALUATION PROBLEMS AND THE MANUAL SECTION 0

PLANNING AND BUDGETING THE EVALUATION SECTION 1

FORMALIZING PROGRAM GOALS SECTION 2

DESCRIBING THE PROGRAM SECTION 3

CHOOSING AN EVALUATION DESIGN SECTION 4

DESCRIBING STUDENT SKILLS FOR SELECTION AND DATA ANALYSIS SECTION 5

SELECTING TESTS SECTION 6

COLLECTING DATA SECTION 7

ANALYZING THE DATA AND REPORTING THE RESULTS SECTION 8

APPENDICES

CONTENTS


| | Page |
|---|---|
| 0. AN ORIENTATION TO THE EVALUATION PROBLEMS AND THE MANUAL | 1 |
| Introduction to the Manual | 1 |
| A Closer Look at the Evaluation Questions | 3 |
| The Problems in Answering the Questions | 6 |
| Popular "Solutions" That Do Not Work | 9 |
| The Positive Side | 11 |
| Contents of the Manual in Brief | 12 |
| 1. PLANNING AND BUDGETING THE EVALUATION | 15 |
| 1-A. A Quick Estimate of Evaluation Costs (Worksheet) | 19 |
| 2. FORMALIZING PROGRAM GOALS | 23 |
| 2-A. Writing Measurable Goals for Student Achievement (Worksheet) | 27 |
| 2-B. Potential Benefits of Bilingual Education Programs (Checklist) | 31 |
| 3. DESCRIBING THE PROGRAM | 35 |
| 3-A. Model for Bilingual Project Description | 39 |
| 3-B. Classroom Observation Guide | 45 |
| 3-C. Category System for Describing Reading Treatment in Bilingual Projects | 51 |
| 3-D. Format for Reporting Instructional Treatment | 55 |
| 4. CHOOSING AN EVALUATION DESIGN | 59 |
| 4-A. A Guide to Evaluation Designs | 63 |
| 5. DESCRIBING STUDENT SKILLS FOR SELECTION AND DATA ANALYSIS | 67 |
| 5-A. Demographic and Biographic Information Worksheet | 71 |
| 6. SELECTING TESTS | 77 |
| 6-A. Selecting an Achievement Test | 81 |
| 6-B. Selecting a Language Proficiency Test | 101 |
| 6-C. A System for Comparing Curriculum Content with the Content of CTBS Spanish and English, Forms B and C | 117 |
| 7. COLLECTING DATA | 149 |
| 7-A. Data Collection Procedures (Checklist) | 153 |
| 7-B. Data Recording Forms | 155 |
| 8. ANALYZING THE DATA AND REPORTING THE RESULTS | 161 |
| 8-A. Data Analysis Checklist | 165 |
| 8-B. Report-Writing Checklist for Bilingual-Program Evaluators | 167 |
| 8-C. Sample Data-Reporting Tables | 175 |
| APPENDIX A: HOW BIG ARE ACHIEVEMENT GAINS? | 179 |
| APPENDIX B: GUIDELINES FOR OTHER EVALUATION AREAS | 193 |
| B-1. Evaluation of Affective Impacts | 197 |
| B-2. Evaluation of Staff Development Activities | 211 |
| B-3. Evaluation of Parent/Community Involvement | 219 |
| REFERENCES | 227 |

# AN ORIENTATION
# TO
# THE EVALUATION PROBLEMS AND THE MANUAL

## Introduction

### An Unconventional Manual

This is not a basic guide to achievement evaluation. It is an attempt to fill in the gaps between the evaluation theory that all evaluators know and the realities of evaluating achievement in bilingual programs. Many of the problems addressed in the manual apply to evaluations of all types of programs but, in most cases, we have attempted to relate them to bilingual programs.

### The Perspective of the Authors

The authors of this manual have reviewed numerous evaluation reports from many types of programs, and have found that most fail to provide convincing evidence of program impact or even to provide interpretable information on student achievement levels. Many reports paint overly rosy pictures of general program effectiveness, while failing to isolate specific, positive impacts that may actually have occurred.

While recognizing the many problems involved in evaluating bilingual programs, and the time and financial constraints under which most evaluators must work, we believe that far more meaningful achievement evaluations are possible in most school districts. This prototype manual is the result of attempting to develop such evaluations in 19 school districts over a two-year period. It is organized around the major, practical obstacles encountered by these districts and provides specific recommendations on how to deal with the problems. The intention is to describe the best available approaches for those districts that have both the commitment to meaningful achievement evaluation and the resources that are required, and to dissuade districts that lack the commitment or resources from expending their efforts on the collection of misleading information.

### The Intended Audience

This orientation section is intended for the local project director or for the federal, state, or local administrator who wants a brief overview of what is currently possible and what is not possible in bilingual-program achievement evaluation. It should also orient the professional evaluator as to where the authors stand on some of the controversial, technical issues in this field.

The body of the manual is intended for the project director and evaluator who must work together, plan, and implement an evaluation. In general, it assumes that the evaluator is familiar with bilingual education and with basic principles of statistics and evaluation design.

## The Evaluation Questions Addressed

This draft manual is not a complete guide to bilingual-program evaluation. It does not yet include such important topics as process evaluation or mastery testing. It focuses only on the evaluation of student achievement and, in particular, on two kinds of questions:

- What is the level of student performance relative to national norm groups or other groups of interest?
- What is the impact of the bilingual program on student achievement compared to other local instruction, past or present?

## The Contents and Organization of the Manual

This is an unconventional guide to evaluating bilingual programs. It deals in depth only with selected, key problems that are either unsolved or widely overlooked in current evaluations. Where some manuals emphasize evaluation principles that are frequently impossible to apply in practice (leaving the decisions and the mistakes to the local evaluator), this manual recommends specific, practicable (though often imperfect) solutions.

Contents. Following this orientation section, the body of the manual is divided into sections according to eight problem areas:

- Planning and budgeting the evaluation
- Formalizing program goals
- Describing the treatment
- Choosing evaluation designs
- Selecting and describing students
- Selecting tests
- Collecting data
- Analyzing the data and reporting the results

In addition, Appendix A contrasts the sizes of errors and effects in educational evaluations, and Appendix B includes sections on evaluating student attitudes, staff development, and parent/community component. These topics would form major sections in a more comprehensive, program evaluation manual.

Organization. Each section is introduced by a two-page overview identifying the one or two key problems in the area. In most overviews, several additional problems are listed for reference with little or no discussion. The eight, two-page overviews constitute the "discussion" sections of the manual.

The remainder of each section is a collection of separate checklists, worksheets, and other items intended to help the project director or evaluator solve the key problems.

The section overviews are relevant to both project directors and evaluators. The various items included in each section may be more relevant to one or the other.

## A Closer Look at the Evaluation Questions Addressed in this Manual

### A Typical "Question-Free" Approach

An evaluation must be planned to answer specific questions if it is to be of any value. Too often, bilingual-program evaluations are not designed to answer any clearly defined question. They are done because they are required by the funding agency, or because someone at the local level believes that evaluation, in some general sense, is important. Frequently, the subject matter to be evaluated and the tests to be used are determined by the test scores available from the district-wide testing program. Everyone knows that "objectives" are important, so an arbitrary gain in raw-score points is chosen. The results, probably positive, are described in an annual evaluation report and may be accepted by the readers as evidence of program effectiveness. In fact, such results mean nothing at all.

Many evaluations add such refinements as comparison groups (which usually turn out to be non-comparable) and language testing (which is ignored in analyzing the reading and math scores). These refinements may provide the raw material for answering some interesting questions, but unless the questions are clearly defined in the minds of the evaluators, they usually are not answered to the satisfaction of the critical reader.

### Different Questions That May Be Asked

There are many different evaluation questions and sub-questions that may be asked and they overlap and interact in complex ways. Evaluation experts generate heated debates over which questions should be asked. The body of this manual addresses two kinds of student achievement outcome questions--(a) the performance level of students relative to other groups of interest, and (b) the impact of a particular program relative to other programs or instructional treatments of interest.

These questions were selected because they are of central interest for many decision makers and because the procedures for answering them are closely related. Other important achievement questions include--(a) whether program objectives were met and (b) whether student skills are adequate for higher education, jobs, or general survival in society. There are also questions of student attitudes, staff development, and parent/community involvement, and of course there are the major areas of cost and process evaluations. All of these questions should be considered in planning a complete bilingual program evaluation. However, only the performance-level and impact questions are covered in Sections 1-8 of this manual.

Student performance-level questions. Establishing relative performance levels is important because national and local comparison groups, where available, can provide realistic standards for program results. Parents, school boards, district administrators, and bilingual-program

staff may all wish to know how program students compare in reading, language, math, and other subjects to:

- the national average (for bilingual program students and non-bilingual program students)
- state and local averages
- other students in the same school
- bilingual program students in previous years.

In order to get meaningful answers to such questions, it is necessary to have: (a) accurate measures of performance for both program and comparison students, plus (b) a clear understanding of the similarities and differences between the program and comparison students. Answering these questions requires careful planning and implementation of the evaluation, but is usually possible in most school districts for English-language subjects. It is usually impossible for native-language instruction.

Program impact questions. Performance-level questions are an attempt to determine whether students are doing well or not in some absolute sense. Impact questions ask whether performance levels are due to the program. In other words--is the program "effective?" Explicitly or implicitly, this question underlies most evaluation designs. It is a question of great interest to local and funding-agency decision makers. However, it is an extremely difficult and expensive question to answer, and very few evaluations succeed in providing convincing evidence on the presence or absence of program impacts.

A major complication in defining the impact question concerns the concept of a "program." If we are looking for effective programs with the intention of spreading them throughout the district (or perhaps to other districts), then the "program" includes only the plans, procedures, and materials that can be exported. Research suggests, however, that the personnel are usually more important than procedures and materials except in a few highly structured programs such as Distar. The distinction between the effects of staff and the effects of program procedures and materials is very important to decision makers and project directors who are trying to improve student learning. It is especially important to know that a clearly effective program from one school may not work well in a new setting.

The other conceptual issue concerns the standard of comparison for the program impact. Program impact is generally measured against a "no-treatment expectation," that is, an estimate of how students would have performed without the new special program. In the case of bilingual-program students, of course, there is always some treatment, whether it is the regular, all-English, classroom program or an alternative special program. It may well be of interest to know how the same children would have fared under these alternatives. However, in this manual, unless otherwise stated, program impact will refer to the situation in which a new program is installed and the question is whether student performance

improves as a result of the program relative to what was being received prior to it.

The difficulty in actually measuring program impacts lies in the small size of the impacts as compared to the relatively large amounts of error and uncertainty in educational evaluations (see Appendix A). In most school districts, it is simply not feasible to expend the resources it would take to isolate the impacts from the sources of error. In general, it would be necessary to start a very elaborate process of collecting baseline data at least two years before the start of the new program. The baseline data would have to include both potential project students and a variety of other students in the district. Then performance data would have to be collected for several years on all students. Program impact would show up as a change in program-student scores relative to the scores of all those students not in the program. In districts that maintain the required data base on a routine basis, impact evaluation should be quite simple. Where an evaluation begins at the same time as a new program, it is <u>usually impossible to demonstrate impact conclusively</u>.

## The Problems in Answering the Questions

### The Five Major Pitfalls

In practice, there are five major problem areas that destroy the credibility of evaluations of all types of programs. These problem areas are:

Undefined program goals and objectives. Much has been written about exactly how to set goals and define objectives. However, we are only concerned here that some reasonable statement of "what-is-taught-when" is available to the evaluation planners. Otherwise, especially in bilingual programs, students may be tested in languages they do not know or on subjects they have not encountered. The problem of goals that are not specified (or are not utilized in planning and reporting the evaluation) is widespread. It applies equally to performance-level and impact evaluations. While conflicting legal, political, and social influences make the program planner's task difficult, there are no insurmountable technical problems in spelling out what is being taught. Project directors and evaluators simply fail to do it.

Inappropriate tests. Obtaining appropriate tests is a serious problem in performance-level evaluations. Language proficiency tests are the subject of considerable controversy among the experts. Theoretical problems can be found with all language tests, and many are difficult and costly to administer. Standardized math computational tests may provide acceptable measures if instructions are translated where required (note also the suitability of the norms, discussed under "comparison groups," below). Standardized English reading tests may also be acceptable, but interpretation is unclear when students have learned to read in another language first. For example, a first-grade English reading test is probably not appropriate for a fourth-grade student who has just begun to learn English but reads fluently in another language. Reading tests are simply unavailable for many other languages. Using test levels that are too difficult or too easy often distorts evaluation results.

For impact evaluations, the testing problems are compounded, because impact evaluations involve the comparison of two or more different instructional treatments. In this situation, any given test will probably favor one treatment over the other simply because it matches the instructional content better. We can speculate that changing tests might even reverse the conclusions as to which treatment is better but, the fact is that no one really knows.

In theory, the testing problems can be solved through further test development but, as a practical matter, the evaluator cannot now find a completely satisfactory set of tests. The problem is further complicated because, in many districts, evaluators are restricted to using tests from the district-wide program.

Lack of comparison groups. For performance-level evaluations, national norms are available for English reading and math, although their relevance to a particular group of bilingual-program students may be open to question. Relevant norms are not widely available for English language proficiency tests, although most language proficiency tests have some form of built-in standards. Local comparison groups obviously exist, but even where they are clearly of interest, it may be difficult or impossible for administrative reasons to obtain the necessary test scores and other descriptive information.

Historically, impact evaluations have implied the need for randomized control groups. In most cases, bilingual-program regulations effectively prohibit such groups. In any case, it now appears that, in a school setting, assignment to a control group may constitute a negative treatment and thus create an inappropriate comparison. Impact studies probably require careful collection of baseline data for several years prior to the start of a new program. For all practical purposes, this rules out any form of precise impact evaluation except in districts that collect these data as a routine matter.

Careless testing procedures. The need for careful testing procedures applies to all forms of evaluation and is well understood by virtually all evaluators. With the possible exception of language testing, there are few technical problems involved. However, a great deal of effort is involved in proper testing, and it is often difficult to focus this amount of effort on testing activities. The result is that questionable testing procedures jeopardize the credibility of many evaluations.

Improper analysis and inadequate reporting. Data analysis and reporting are the final links in the evaluation process. Like test administration, data analysis and reporting present no major technical problems. All that is required is the thoughtful application of existing evaluation methodology and the careful documentation of what was done. Nevertheless, few evaluation reports are even marginally adequate. In general, the evaluation questions are not clearly defined, information on the tests and testing procedures is incomplete, inappropriate evaluation models are applied, and little or no attempt is made to tie results to program features. Of all the major pitfalls, however, this one may be the easiest to eliminate. Given adequate incentives and guidelines, most evaluators could produce satisfactory analyses and reports.

## Additional Obstacles Related to Program Characteristics

In addition to the five pitfalls above, which apply to evaluations of all special programs, there are four widespread problems that are of concern in many bilingual program evaluations:

Evaluation of fragments of a program. Most bilingual programs are designed to cover several grades, and many important skills are not even introduced at the lowest grades. Such programs are often started at the lower grade levels and expanded upward, one grade per year. A K-6 program cannot be evaluated by observing one or two of the lower grades. A long-term, longitudinal study is required, but such studies present many problems. In fact, student turnover makes most program evaluations longitudinal in theory only.

Evaluation of new programs. Bilingual education is characterized by new and constantly evolving programs. There is a great deal of pressure to provide immediate evidence of positive results, but there is simply no way to do a meaningful outcome evaluation of a program that is only partially in place or is in a state of flux.

Variation in instructional treatment. Treatment in bilingual programs often varies widely among students, even within a single classroom. Meaningful evaluation requires a clear understanding of what happens to each student, but when instruction is described clearly, it may become obvious that only a few students received any one treatment. The different groups may be too dissimilar to aggregate but too small to analyze separately.

Testing of young children. The testing of young children, especially those below the third grade, is notoriously difficult. Many bilingual programs, however, focus heavily on the lowest grades. There is no obvious answer to this problem.

## Evaluator Availability and Turnover

An adequate evaluation requires a lot of evaluator time and careful adherence to a long-range plan. These requirements are difficult to meet, especially in small districts. Frequently, the evaluator has very limited time and resources; when he or she leaves to take a new job, the work of several years may be lost. While these are management problems rather than technical ones, they are major causes of inadequate program evaluations.

## Popular "Solutions" That Do Not Work

The frustrations generated by the kinds of problems described above have lead to many misguided attempts at solutions. Some fail to answer the impact question, but do answer other questions of possible interest. Others are of no use at all.

### Approaches That Should Never Be Used

Posttest minus pretest. In lieu of any better ideas, many evaluators simply subtract pretest scores from posttest scores and compute the significance of the difference. Since almost all groups of children make some gains, even when they are falling rapidly behind their peers, this approach is of no value at all. A popular variation, selecting a gain of some arbitrary number of raw-score points as the program target, is no improvement.

Grade-equivalent scores (the month-for-month-gain myth). Analyses based on grade-equivalent scores still, unfortunately, appear all too frequently. They are based on the mistaken belief that a gain in test scores of one or more months for each month of instruction represents good progress. This is not true. Grade-equivalent scores provide an illusion of simplicity but, in fact, they are almost impossible to interpret, even for specialists in test construction. Grade-equivalent scores should never be used by anyone for any purpose whatsoever.

IQ-based formulas. From time to time an attempt to use IQ scores appears as the basis for evaluating reading or math performance. The idea that IQ tests provide an absolute standard against which to compare a specific skill is simply a misunderstanding. IQ-based formulas are not appropriate for use in bilingual program evaluations.

Subjective data. As a last resort, evaluators sometimes fall back on subjective data, usually teacher reports. While such reports are always useful in interpreting results, they can never be assumed to represent reliable, valid measures of student performance.

### Approaches That Are Widely Misused

Criterion-referenced testing. A great deal of mistique has been generated around criterion-referenced tests, and some evaluators suggest that they solve the major problems faced by evaluators. Actually, what the criterion-referenced-test advocates have done is to change the question that is being asked. Criterion-referenced tests may provide information as to whether program objectives have been met (although those objectives may be quite arbitrary). Measuring performance level or program impact still requires reliable, valid tests with adequate range of difficulty (no floor or ceiling effects). In principle, criterion-referenced tests could meet these requirements but, in practice, most do not.

Gap-reduction models. "Gap reduction" is a term that appears in the bilingual program evaluation literature. It usually means either (a) students get closer to the national norms, or (b) students get closer to some dissimilar comparison group. The former is simply a special case of the norm-referenced model, which is useful for performance-level evaluation but generally not for program-impact evaluation. The latter is an example of non-random comparison groups (see below). The important point is that "gap-reduction" is simply a new name for familiar designs. The new name does not change their strengths or weaknesses.

Non-random comparison groups. Many bilingual program evaluations make use of non-random comparison groups, that is, different kinds of students who are receiving different instructional treatments. As part of a performance-level evaluation, such comparisons may be of great interest to local decision makers and program staff. In general, however, such comparisons do not by themselves provide program impact information because student differences are confounded with program differences.

Time-series or historical data designs. In combination with non-random comparison groups, time-series designs may provide the most accurate program-impact evaluations. However, the baseline (pre-program) data must be collected very carefully. Long-term planning is usually required well before the start of a new program, and this presents a serious practical problem. For example, none of the 19 sites in the bilingual-PIP field test had the data on hand to enable them to use this design.

## The Positive Side

The preceding pages have painted a very bleak picture of the current state of achievement evaluation. At this point, you may be asking if things can really be this bad and, if so, whether there is any point in ever trying to evaluate student achievement. The answers are (a) that the current situation is definitely as bad as we have painted it, but (b) that useful evaluations are quite possible in almost every school district. However, there are two important qualifications to the latter answer:

- You must be clear about the questions you are trying to answer, and it may not be feasible in your district to answer every question that you would like to answer. Requirements for answering two kinds of questions are summarized on this page.

- You must attend to all of the requirements for carrying out a useful achievement evaluation. These requirements are organized under eight section headings in this meanual. The section contents are summarized on the following two pages.

### The Questions of Program Impact on Achievement

These questions are very difficult and costly to answer in most districts. It is not overly difficult to determine the performance level of the students (see below), but it is very difficult to determine how much a particular program has contributed to the performance level. For example, to determine the impact of a new program on achievement, you must have accurate data on both program and non-program students, both before and after the introduction of the program. If the program is the only change in the district, and if the achievement of program students improves in relation to the non-program students, then you can probably conclude that the program produced the improvement.

### The Questions of Student Performance Levels

It is quite possible to get good measures of student performance levels, simply by following the procedures described in this manual. This information can tell you where program students stand in relation to other students in the district or in publishers' norm groups. It will also let you compare the achievement levels of program students from year to year. It will probably not tell you with any certainty whether the program was the cause of improvements or declines, but it may alert you to problems or support your judgments about program effects.

### Other Questions

Of course there are other important evaluation questions that should be answered, and some of them are less difficult to answer than are the achievement questions. These would include whether the program is operating as planned (process evaluation) and whether the students are achieving specific, instructional objectives. A complete evaluation addresses all of these questions.

## Contents of the Manual in Brief

1. Planning and budgeting the evaluation. Evaluating a bilingual program involves developing a design, selecting tests, training testers, supervising data collection, analyzing data, writing reports and presenting findings to district personnel and program staff. Some evaluators are also asked to train staff in diagnostic procedures. Few budgets permit evaluators to work in any depth on the basic problems that exist in each of these areas (see below). Many evaluators have too little time for even a minimal effort on the complete set of tasks listed above. This section provides rough guidelines as to what can be accomplished at different levels of effort.

2. Formalizing program goals. Few programs have written down exactly what they hope to accomplish (with rationales explaining why their instructional approach should succeed) in each subject area and grade level. This section describes the kinds of brief, clear goals that are essential for selecting tests and interpreting evaluation results. Rationales are treated in the next section.

3. Describing the treatment. In order to interpret results, the evaluator and the audience for the report must know how much time the students spent on each topic, and exactly how that time was spent. Reports often include discussions of gains (or losses) for students who actually received no special training in relevant areas. This section provides a detailed list of program features that should be considered by the evaluator and summarized in the evaluation report.

4. Choosing experimental designs. An experimental design is a formal statement of a question or set of questions that the evaluation is intended to answer. As discussed above, the most important feature of a design for evaluating a bilingual project is the control group or other standard of comparison. It was pointed out that there usually is no practicable way to get a precise answer to the basic question of whether students do better in a program than they would have done without it. However, important questions such as whether students are doing "well enough," "better than last year," "better than students in other projects," and so on, can often be answered precisely enough for practical purposes. This section describes the basic questions that can be answered, and the designs that will answer them.

5. Selecting and describing students. The ways in which students are selected influence evaluations because they determine the characteristics of the program groups and the availability of comparison groups. In addition, descriptions of student backgrounds and educational experiences are needed in order to interpret the results of the program evaluation. This section explains some of the ways in which student selection approaches affect impact evaluation designs and indicates the information required for each student in order to understand her or his performance in the program.

6. Selecting tests. The particular tests used to measure language, reading, math, and perhaps other skills are of extreme importance if an evaluation is to povide any useful information at all. A great deal has been written about the principles to follow in choosing tests, and these principles must be observed. In practice, however, there are very few acceptable tests to choose among, and the most important thing is to pick a test that is relevant to what is taught in the program. This section lists the basic features to look for, names the most widely used tests, and provides an example of a detailed analysis of test content.

7. Collecting data. Experience in Title I and other evaluations has shown that the size of program impacts on student scores may be as little as a few raw-score points, even in a very good program. This may also be the case in bilingual education programs. Major violations of correct testing and scoring procedures appear to be very commmon in school settings and may often be the real sources of apparent program successes or failures. This section reprints a list of basic testing and scoring rules that must be followed for a meaningful evaluation.

8. Analyzing the data and reporting the results. Convincing evaluation reports from bilingual programs (or, for that matter, from any educational programs) are almost non-existent. In general, appropriate data analysis consists of determining if there appear to be program impacts and, if so, exactly what causes the apparent impacts. Only the simplest of statistical techniques are required in many cases, although in some evaluations, multiple regression techniques may be required to adjust for students' socioeconomic status, educational background, and other relevant factors. Careful, common-sense detective work is always required in locating causes. This section suggests which analyses to conduct and what to include in the report. References to more detailed treatments of data analysis procedures are also included.

SECTION 1

# PLANNING AND BUDGETING THE EVALUATION

## Major Content Item

1-A A QUICK ESTIMATE OF EVALUATION COSTS (WORKSHEET)

## 1. PLANNING AND BUDGETING THE EVALUATION

An evaluation design for a bilingual program may be as simple as an outcome evaluation to meet minimum Title VII requirements or as complex as a complete process and outcome feedback system for teachers, project directors and district personnel. We assume here that a decision has been made as to the general nature of the evaluation and that some performance-level and/or impact evaluation will be included.

One of the first planning steps for the project director is to check on available evaluation resources in the district and, assuming she or he has the authority, to decide whether to use an in-house evaluator or hire someone from outside. The possibility of obtaining special skills or facilities from a university or private evaluation specialist must be weighed against the potentially lower cost to the project budget and more convenient working relationship with an evaluator on the district payroll.

Key Problem: Specifying Evaluation Tasks

Perhaps the major evaluation planning problem encountered in the PIP field test was the discrepancy between (a) the many evaluation tasks that needed doing, and (b) the time and resources available to do the job. Typically, evaluators are expected to help in designing the evaluation, and selecting tests, and have full responsibility for training testers, supervising testing, analyzing data, writing reports and presenting results to district personnel. Many are also involved in process evaluations and in providing feedback to teachers. Budgets for such services may cover less than 25 days of evaluator time, far too little for anything more than a superficial effort. While each district must decide on its own evaluation needs, this section provides some rough guidelines as to how much services will cost from external evaluators.

## Related Issues

Long-range planning. Evalution of bilingual programs requires long-term, longitudinal evaluation designs. The evaluations should be planned and budgeted on this basis.

Evaluator attrition. From time to time, evaluators resign and must be replaced. Occasionally this happens in the middle of the year. Unless all evaluator activities are carefully documented, or someone who remains with the project is thoroughly familiar with what has been done, you can expect to loose much of the evaluation for the year. This possibility must be weighed against the substantial costs of documenting every step or involving a second person in the evaluation.

Conforming to district policies. In many districts, test selection and other key evaluation decisions are constrained by district policies, and the credibility of bilingual program evaluation may suffer. The project director and evaluator must use good judgment as to whether the potential improvement to the evaluation justifies the problems in deviating from district practices. Where deviations are not justified or are simply not possible, all that can be done is to document the effects on the evaluation in the evaluation report.

## A Quick Estimate of Evaluation Costs

Evaluation costs will vary widely from district to district. This worksheet is intended to give a quick estimate of the cost of an achievement evaluation in a bilingual program, using an external evaluator. It may be possible to obtain more time from a district evaluator at less cost to the project, depending on local policies and resources.

Two levels of effort are given. The "minimum" level is included as a lower bound on costs. This level is not uncommon in school districts and may meet evaluation requirements of some funding agencies, but it will not provide useful information in most cases. The "major" level represents a more realistic estimate for an adequate evaluation. The worksheet is only intended to provide a ballpark estimate for a few minutes work. It cannot substitute for careful evaluation planning.

### Evaluation Tasks Not Included

The worksheet addresses only a limited part of a complete evaluation. Do not forget to include the following additional tasks in your complete evaluation plan.

1. Needs assessment
2. Process (formative) evaluation
3. Teacher workshops on evaluation
4. Student diagnostic testing
5. Cost analysis of program

### Evaluator Rates

Use actual rates if you know them. Otherwise, select one of the following for a rough (1980) estimate.

| | | |
|---|---|---|
| 1. | Qualified independent evaluator (no overhead) | $100+ per day |
| 2. | Evaluator contracted through an evaluation company (includes overhead) | $200+ per day |
| 3. | Senior evaluator from major educational research company (includes overhead) | $300+ per day |

## Cost-Estimate Worksheet

### Typical Level of Effort Per Task

1. Evaluation planning (produce written plan) ____ days

   Minimum (Routine performance-level evaluation in a small, familiar program): 3 days
   Major (Comprehensive impact evaluation in a large, unfamiliar program): 10+ days

2. Select tests ____ days

   Minimum (Familiarization with district tests) 1 day
   Major (Review commercial achievement tests and match to curriculum, review language and affective tests, develop staff development and parent/community instruments, document process): 20+ days
   Develop achievement or language tests: Not feasible

3. Train testers ____ days

   Minimum (Small project, experienced testers, pre- and posttesting): 2 days
   Major (Large project, inexperienced testers, achievement and language tests, pre- and posttesting): 6 days

4. Supervise testing ____ days

   Minimum (One day each, pre- and posttest): 2 days
   Major (monitor all testing): 14+ days

5. Conduct classroom visitations ____ days
   Minimum (one visit pr classroom during the course of the school year; small program; 6 classrooms, 2 classes visited per day): 3 days
   Major (two visits per classroom; large program; 24 classrooms, 2 classes visited per day): 24 days

6. Analyze data ____ days

   Minimum (Prepare achievement data for standard computer analysis, small program, pre- and posttest): 5 days
   Major (Comprehensive evaluation, large program, thorough study of computer print outs and matching results curriculum, pre- and post-test): 20+ days

7. Write reports ____ days

| | |
|---|---|
| Minimum (One report, routine evaluation): | 5 days |
| Major (Pre- and posttest reports, comprehensive evaluation, polished posttest report): | 25+ days |

8. Present findings ____ days

| | |
|---|---|
| Minimum (Discuss with project director, pre- and posttest): | 2 days |
| Major (Formal presentation for school board, discuss with project direcor, feedback to teachers): | 10+ days |

Total number of evaluator days (20 to 105+) ____
x Cost per day ____
Total evaluator cost ____

Additional-Cost Items

| | | | |
|---|---|---|---|
| 1. | Test administrators (if needed) | local substitute rate | _____ |
| 2. | Tests | $1.00 per student* | _____ |
| 3. | Test scoring (by hand or by scoring service) | $ .70 per student* | _____ |
| 4. | Computer time for analysis | less than $200* | _____ |
| 5. | Secretary time | local rates | _____ |
| 6. | Printing costs for reports | local rates | _____ |

*May vary considerably. Use for rough estimates only.

SECTION 2

# FORMALIZING PROGRAM GOALS

MAJOR CONTENT ITEMS

2-A. WRITING MEASURABLE GOALS FOR STUDENT ACHIEVEMENT (WORKSHEET)

2-B. POTENTIAL BENEFITS OF BILINGUAL EDUCATION PROGRAMS (CHECKLIST)

## 2. FORMALIZING PROGRAM GOALS

This section is concerned with the very basic and widely ignored task of spelling out the general kinds of impacts that a particular bilingual program is expected to make. (Section 3 deals with describing how the program is expected to meet these goals.) An absolute minimum for a meaningful evaluation design should include:

- Student achievement goals
- Student affective goals
- Parent/community goals
- Staff development goals

For each set of goals, implementation time schedules are necessary in order to determine what can be evaluated the first year of the program, the second year, and so on. Ideally, each set of goals should also be discussed in relation to a needs assessment, in order to justify the goals and demonstrate that they are neither trivial nor unrealistically difficult.

This section focuses on student achievement goals.

Key Problem: Defining Student Achievement-Goal Categories

While there are many important considerations in specifying goals, this section is limited to one that is absolutely essential. Goals must, at the very least, be broken down by:

a. Subject areas (e.g., reading, language, math)
b. Languages to be used (e.g., English, Spanish, etc.)
c. Student language proficiency category (e.g., English: limited or proficient, Spanish: limited or proficient)

Using these examples, the first categories of goals would be those for

- Reading, in English, for fluent-English students.
- Reading, in English, for limited-English students.

This is a most rudimentary breakdown, but it will be noted that it requires 3 x 2 x 2 = 12 sets of goals, and these goals must be completed for each grade level.

This section includes a worksheet and a checklist to help in specifying goals. However the first step must be to define categories as in this example. Few districts provide even this basic breakdown, but without it, the evaluation means little.

## Related Issues

Legal requirements. The goals of most programs must meet local, state, and possibly federal guidelines in addition to those developed by project personnel. Compliance with these guidelines may be the major consideration in the goals you state and the way you state them but, in most cases, it should still be possible to include all of the basic categories of goals (indicated under "Key Problems") within the legal constraints.

Short-term and long-term goals. Many projects fail to distinguish between long range goals that can only be evaluated over a period of several years, and short-term, intermediate goals that may be relevant to a one-year evaluation. This is an especially important problem in bilingual programs since (a) some long-term goals (e.g., improved English skills when compared to a non-bilingual classroom) may not apply until the later grades, and (b) many bilingual programs experience high student turnover. Long-term goals may not apply to short-term project students, and therefore special goals may be required.

Goals for follow-up services. It is widely recognized that students who meet existing criteria and are transferred from a bilingual program to a conventional classroom may still be in need of special follow-up services. In districts that provide such follow-up services, the follow-up goals should be clearly specified and carefully integrated with the bilingual program goals as well as with non-bilingual program goals.

## Worksheet for Writing Measurable Goals for Student Achievement

A Worksheet for Writing Measurable Goals for Student Achievement is included in this chapter. It is designed for use by the evaluator in conjunction with the program staff, and should be helpful in clarifying to the staff how to go about writing goals systematically. The chart is organized by subject area, by language group, and by language of instruction. It can be adapted to meet local needs by adding or deleting categories and including higher grade levels. The category of English reading for Spanish dominant students is filled in to illustrate use of the chart.

## Worksheet for Writing Measurable Goals for Student Achievement

Year _____

| Subject Area and Language Group | Grade Level: Kindergarten Taught? | Kindergarten Mean Score Expected on Measure | 1st Grade Taught? | 1st Grade Mean Score Expected on Measure | 2nd Grade Taught? | 2nd Grade Mean Score Expected on Measure | 3rd Grade Taught? | 3rd Grade Mean Score Expected on Measure |
|---|---|---|---|---|---|---|---|---|
| I. First and Second Language Skills | | | | | | | | |
| A. For Spanish dominant students, limited in English proficiency | | | | | | | | |
| 1. Spanish skills | | | | | | | | |
| a. listening | | | | | | | | |
| b. speaking | | | | | | | | |
| c. reading readiness | | | | | | | | |
| d. reading | | | | | | | | |
| e. writing | | | | | | | | |
| 2. English skills | | | | | | | | |
| a. listening | | | | | | | | |
| b. speaking | | | | | | | | |
| c. reading readiness | | | | | | | | |
| d. reading | no | none | no | none | 5 kids* / 40 kids | none | 38 kids / 40 kids | mean of 35%ile on CTBS |
| e. writing | | | | | | | | |

*5 of the 40 limited English students received instruction.

## Worksheet for Writing Measurable Goals for Student Achievement

Year _____

| Subject Area and Language Group | Grade Level: Kindergarten Taught? | Kindergarten Mean Score Expected on Measure | 1st Grade Taught? | 1st Grade Mean Score Expected on Measure | 2nd Grade Taught? | 2nd Grade Mean Score Expected on Measure | 3rd Grade Taught? | 3rd Grade Mean Score Expected on Measure |
|---|---|---|---|---|---|---|---|---|
| B. For English dominant students, fluent in English. | | | | | | | | |
| 1. Spanish skills | | | | | | | | |
| a. listening | | | | | | | | |
| b. speaking | | | | | | | | |
| c. reading readiness | | | | | | | | |
| d. reading | | | | | | | | |
| e. writing | | | | | | | | |
| 2. English skills | | | | | | | | |
| a. listening | | | | | | | | |
| b. speaking | | | | | | | | |
| c. reading readiness | | | | | | | | |
| d. reading | | | | | | | | |
| e. writing | | | | | | | | |

## Potential Benefits of Bilingual Education Programs

### Introduction

One problem in the evaluation of bilingual projects is that the expected benefits of a program are broad, and yet most evaluations tend to focus only on those areas of student achievement for which tests are readily available. Many other important outcomes for students, as well as for the school and the community, may be overlooked. For example, one of the immediate strong impacts of a well implemented bilingual program on limited English-speaking students is a sense of belonging and the ability to participate in academic and social activities. This list of potential benefits is designed to be used by evaluators in conjunction with program staff. The list can be used in setting and prioritizing goals, and selecting which ones will be documented or measured. It can also be used to highlight unintended outcomes of the program.

### Instructions

For the first column, "Intended Result of Project," check those items that are explicit goals of the program. For the second column, "This is Being Measured," check those items that you are measuring or documenting. For the third column, "We See This Happening," check those items that you feel are occurring as a result of the bilingual program.

## Potential Benefits of Bilingual Education Programs

District _____

Project Director _____

Evaluator _____

Date _____

| Intended Result of Project | This Is Being Measured | We See This Happening |
|---|---|---|
| | | |

### Students

1. More meaningful education for students of limited English proficiency since students are able to participate fully in their educational experiences
   - student and teacher can better communicate with one another
   - student is able to participate in broader range of school activities including social as well as academic activities
   - student is better able to relate to and profit from instructional materials
2. Increased verbal expression
   - increased use of native language
   - increased use of second language
3. Greater sense of belonging due to acceptance of language and cultural diversity
4. Increased benefit from teacher guidance and counseling due to use of native language of student
5. Reduced alienation between parents and children because of school's inculcation of respect for student's home language and culture

## Potential Benefits of Bilingual Education Programs (Continued)

| | Intended Result of Project | This Is Being Measured | We See This Happening |
|---|---|---|---|
| **Students** (continued) | | | |
| 6. Improved attitude toward school | | | |
| 7. Improved attitude toward certain school subjects | | | |
| 8. Improved self-concept | | | |
| 9. Improved motivation | | | |
| 10. Better race relations | | | |
| • increased interethnic play at school and at home | | | |
| • improved attitudes toward other ethnic groups | | | |
| • improved attitudes toward other languages | | | |
| • improved attitudes toward other cultures | | | |
| 11. Other _____ | | | |
| **School District** | | | |
| 1. Improved school climate | | | |
| 2. Improved relations among staff | | | |
| 3. Improved community-school relations | | | |
| 4. Greater degree of compliance with legal mandates | | | |
| 5. Decreased district spending due to decreased retention rate | | | |

## Potential Benefits of Bilingual Education Programs (Continued)

| | Intended Result of Project | This Is Being Measured | We See This Happening |
|---|---|---|---|
| **School District** (continued) | | | |
| 6. Higher attendance bringing in higher ADA | | | |
| 7. Reduced adult/student ratios | | | |
| 8. Additional staff | | | |
| 9. Additional training and professional development for instructional staff | | | |
| 10. Additional materials, facilities, equipment, supplies | | | |
| 11. Other _____ | | | |
| **Parents and Community** | | | |
| 1. Greater participation in school activities | | | |
| 2. Increased knowledge of bilingual program operation | | | |
| 3. Improved race relations | | | |
| 4. Improved community-school relations | | | |
| 5. A more informed citizenry (resulting from parent, adult education programs) | | | |
| 6. Additional jobs for the community | | | |
| 7. Other _____ | | | |

# DESCRIBING THE PROGRAM

SECTION 3

Major Content Items

3-A. MODEL FOR BILINGUAL PROJECT DESCRIPTION

3-B. CLASSROOM OBSERVATION GUIDE

3-C. CATEGORY SYSTEM FOR DESCRIBING READING TREATMENT IN BILINGUAL PROJECTS

3-D. FORMAT FOR REPORTING INSTRUCTIONAL TREATMENT

## 3. DESCRIBING THE PROGRAM

Accurate impact measures represent only one side of the task of producing a meaningful, useful impact evaluation. The other side consists of the description of the program that produces the impact and the analysis of why the program treatment leads to (or fails to lead to) desired impacts. At the stage of planning an evaluation, every program goal should be compared with the program description to be sure that there is enough reason to expect the goal to be met to justify the effort and expense of the evaluation. If, for example, a new bilingual program replaces all first-grade English reading with Spanish reading, there is no reason to expect the program to make dramatic improvements in first-grade, English reading scores. At the report-writing stage, the link between program features and impacts must be made perfectly clear to the reader. These obvious principles are almost universally ignored in evaluations, and bilingual program evaluations are no exceptions.

As with the specification of goals, a thorough, detailed description of the bilingual program is highly desirable. However, this section is concerned primarily with the very basic, rudimentary description that is absolutely essential. In addition to features discussed below, this basic description must include (a) the broad context of the school and community, (b) the comparison group (or norm group) treatment, (c) the legal requirements affecting the program, and (d) teacher characteristics. (See also Section 5, Describing the Students.)

Key Problems

Descriptions for the use of the project director and evaluator. The project director and evaluator need to have a very clear picture of the project they are operating and evaluating. At a minimum, they must be able to relate impacts to treatments in enough detail to suggest where changes are needed and what changes to make. The first three items in this section are intended to help in developing the treatment description that they require.

Descriptions for the evaluation report. The knowledgeable readers of the evaluation report will simply not believe accounts of major impacts unless a plausible explanation (i.e., a learning situation substantially different from, and apparently superior to, the conventional classroom) is offered. This description of treatment must be clear, but need not be as detailed as that for the project director and evaluator. (See Item 3-D.)

## Related Issues

Describing variation in treatment. In most programs, the treatment varies for different students depending on their language skills, reading and math skills, and other factors. In such cases each different treatment must be described, and students must be grouped for the data analyses according to the treatment they received.

Longitudinal descriptions of treatments. In describing the bilingual program, it is essential to make clear what the student experiences throughout all of his or her years in the program. Bilingual programs often include a coordinated curriculum for grades K-6, and the complete program must be described.

## Model for Bilingual Project Description

A description of the bilingual project is an essential part of an evaluation report. The Model for Bilingual Project Description presented here was developed during the Bilingual PIP dissemination study and is based on RMC staff expertise and the experience of the bilingual PIP field test. Literature from the field of bilingual education was also examined, and ideas were incorporated from similar models that have been developed (Mackey, 1977) as well as from a wide variety of more general current works (see, for example, Center for Applied Linguistics, 1977, 1978).

The model is divided into three major areas: (1) overview, (2) instruction, and (3) management. Each area consists of lists of categories to be considered in providing a comprehensive project description. Though it is always somewhat artificial to divide an organic whole like a project into a system of categories, the model is intended to be as systematic and comprehensive as possible. The evaluator may utilize those sections of the model that are particularly appropriate to the project being described.

MODEL FOR BILINGUAL PROJECT DESCRIPTION

1. OVERVIEW OF BILINGUAL PROJECT

   1.1 Project Summary

      1.1.1 Major Goals

      1.1.2 Target Student Population

         - Language characteristics
         - Achievement levels

      1.1.3 Grades and Number of Classrooms Served

      1.1.4 Portion of School Day Covered

   1.2 Local Context

      1.2.1 Community Characteristics

         - Languages
         - Ethnicity
         - SES
         - Mobility
         - Size

      1.2.2 LEA Description

         - Size
         - Financial status of district
         - Facilities available for project

      1.2.3 Relevant History of LEA and Community

         - Special projects
         - Desegregation

2. INSTRUCTIONAL APPROACH

   2.1 Content of Instruction

      2.1.1 Content Areas Covered

      2.1.2 What Determines Content

      2.1.3 Other Content Features

         - Relationship of content to goals
         - Articulation of project content with existing district curriculum

## 2.2 Presentation of Content

### 2.2.1 Instructional Models or Theories

- Bilingual education model
- Other model

### 2.2.2 Methodologies for Bilingual Education

- Language of instruction
  - — general language use plan for teacher and student over length of project
  - — daily instructional time in each language
  - — variations for different student groups
  - — criteria for establishing language of instruction
- Approach to non-standard forms
  - — acceptance
  - — form of corrections
- Approach to second language instruction
  - — formal instruction
  - — functional use of second language for content instruction and other activities
- Approach to reading instruction
  - — language in which students learn to read
  - — criteria for beginning reading in second language

### 2.2.3 Specific Methodologies for Each Subject Area

### 2.2.4 Rate

- Variation in pace of instruction for individuals or groups
- Time on task
  - — minutes per day per content area (see Scheduling, 2.4)
  - — proportion of time student is actively engaged in producing responses for which s/he gets feedback

2.2.5 Self-Concept Development and Motivation (Aspects of program that may motivate students and improve their self-concept)

- Appropriate content and language of instruction
  - — using $L_1$ for instruction
  - — accepting the language of the student
  - — content that relates to experience of students
  - — culturally relevant material
- Improved affective climate
  - — placing equal value on both languages and cultures
  - — insuring student success
  - — involving parents
  - — teacher as a role model
- Discipline approach
  - — philosophy
  - — guidelines/control over approach
- Special reward systems
  - — prizes, privileges

2.2.6 Materials

- Core materials in use
  - — commercial
  - — locally developed
- Appropriateness
  - — linguistic
  - — cultural

2.2.7 Personnel Roles in Classroom

- Teachers
- Aides
- Parents
- Peers
- Resource staff

2.3 Student Selection

2.3.1 Entry Criteria and Procedures

2.3.2 Exit Criteria and Procedures

2.4 Scheduling

2.4.1 Grouping and Regrouping

- Across classes
- Within classes

2.4.2 Daily Schedules

3. MANAGEMENT

3.1 Staff Organization

3.1.1 List of Staff Members and Time Commitment

3.1.2 Organizational Structure

3.1.3 Qualifications

3.1.4 Selection Procedures

3.2 Staff Roles (describe responsibilities)

3.2.1 Project Director

- Style of leadership as determined by project and LEA
- Funds and budgets
- Public relations
- Administration
- Overseeing instruction
- Staff training
- Developing and ordering materials and equipment
- Staff recruiting and hiring

3.2.2 Teachers

- Planning instruction
- Implementing instruction
- Non-instructional responsibilities

3.2.3 Aides

3.2.4 Other Staff

- Instructional coordinator
- Community coordinator
- Evaluator

3.3 Staff Development (describe)

3.3.1 Needs Assessment

3.3.2 Structure of Training

- Pre-service
- In-service

3.3.3 Characteristics of Training

- Appropriateness for staff of differing levels of knowledge and experience
- Practicality
- Coordination with degree programs
- Integration with other training

3.3.4 Audiences Trained

- Project staff included
- Inclusion of non-project staff

3.4 Parents and Community

3.4.1 Parent Involvement in School Affairs

3.4.2 Community Input in Program Planning

- Advisory group

3.4.3 Community Support for Project

3.4.4 Parent Education

3.4.5 Parent Conferences/Counseling

3.5 Communication

3.5.1 Staff Relations

3.5.2 Relations with Non-Project Staff

- District administrators
- Principals
- Non-project teachers
- School board

3.5.3 Dissemination of Project Information

- School personnel
- Parents and community

## Classroom Observation Guide

Reference was made earlier to the need to establish the amount of participation students have in the program, as well as the type of instruction received by the participants. The evaluator must know and describe what it is that is being evaluated.

The following form is provided as an example of a classroom analysis tool that can be used for the purpose of documenting the qualitative and quantitative characteristics of features deemed essential to bilingual education. The value of conducting classroom observations will become apparent when attempting to analyze data and when providing feedback to the school. An evaluator, as an observer in a classroom, is there to gauge the potential for certain practices and characteristics previously identified as desirable. For example, it is agreed that to have parents participating in the school is desirable. However, an evaluator may be unlikely to witness such an event during a classroom visit. Therefore, it would be necessary for the evaluator to interview teachers, PAC members, and to see some documentation of parent participation. The Model for Bilingual Project Description can be used to construct the necessary interview guides for any of the program components, and therefore supplements the Classroom Observation Guide.

Classroom observation is a complex undertaking. It should be done as frequently as possible in order to obtain a reliable description. In addition, the person(s) conducting the observations should be adequately trained. This is a simple, rudimentary guide that can be used by districts to develop guides that are more applicable to their own needs.

## CLASSROOM OBSERVATION GUIDE

Date _____

School District _____

School _____

Teacher _____

Aide _____

Grade _____

Observer _____

Duration of Observation _____

Interview Required Yes _____ No _____

## CLASSROOM OBSERVATION GUIDE

Subject: _____ Lesson: _____
(Use one sheet per subject)

Methodology/Theory: _____

Materials: _____

_____

| GROUP | | INSTRUCTOR(S) AND ROLE | LANGUAGE OF INSTRUCTION | | DURATION PER WEEK (in hours of exposure) | | NOTES |
|---|---|---|---|---|---|---|---|
| Characteristics | Size | | Teacher | Student | Subject | Language | |
| | | | | | | | |

BASIS FOR GROUPING:

Criteria: _____

Assessment Method: _____

Permanence: _____

| |
|---|
| **Composition of Classroom:**<br>**Number of students of limited English proficiency _____**<br>**Degree of segregation/integration:** |
| **Physical/Structural Layout:** |
| **Approach to Culture and Heritage in Lessons Observed:** |
| **Visual Displays:** |

| |
|---|
| Nature/Tone of Interaction: |
| Language Use by Teachers and Students in Non-Instructional Settings: |
| Student Attendance, Turnover: |
| Other Observations (Optional): |

# CATEGORY SYSTEM FOR DESCRIBING READING TREATMENT IN BILINGUAL PROJECTS

Interpretation of evaluation data from bilingual projects is not easy due to the complex nature of the projects as well as the variation in instructional treatments across classrooms and within classrooms. The more teachers provide instruction to meet the particular needs of individuals or different groups of students, the more difficult it becomes for the evaluator to document the instructional treatment received by students in the program. Nevertheless, the treatment must be described in order to aggregate and analyze data in a meaningful way and in order to interpret findings adequately.

In the area of reading, for example, it is essential to know how much reading instruction, if any, was received by each student in each language. The reader of an evaluation report should not be lead to assume that a reading test score for a particular student in a particular language represents one full year of reading instruction in that language if such is not the case for all students. By using this category system, inappropriate aggregation and misinterpretation due to lack of information can be avoided.

The Category System can be used to provide a very basic description of the type of reading instruction received by each student. The system was developed based on observations of a large number of bilingual programs. It was then pilot tested in two school districts and refined based on user feedback. The eight categories have been designed to include the most common instructional situations encountered in bilingual projects. They include types of reading instruction for students of limited English proficiency as well as for those proficient in English. They are offered for Spanish/English and French/English programs, but may be adapted to any language. For a more thorough description, the precise amount of time that each type of reading instruction was provided to each student can be recorded.

The assignment of a category label should be done by the teacher providing instruction. This normally requires 20 to 30 minutes per teacher. Considerably more thorough interpretation of program outcomes is possible if this information is recorded for each year of a student's participation in the project.

CATEGORY SYSTEM FOR DESCRIBING READING TREATMENT
IN BILINGUAL PROJECTS
Spanish/English Form

| Category Label | Definition |
|---|---|
| S | Daily reading in Spanish only, for entire year. No English reading. |
| SE | Daily reading in both Spanish and English for entire year. |
| S-SE | Daily reading in Spanish only, from fall to mid-year (sometime between December and March). Daily reading in both Spanish and English from mid-year to end of year. |
| S-E | Daily reading in Spanish only, from fall to mid-year (sometime between December and March). Spanish reading discontinued at mid-year. Transfer to daily reading in English only from mid-year until end of year. |
| E | Daily reading in English only, for entire year. No Spanish reading. |
| E-ES | Daily reading in English only, from fall to mid-year (sometime between December and March). Daily reading in both English and Spanish from mid-year to end of year. |
| ? | Reading treatment unknown. |
| O | Other. Please describe. _____ |

Instructions

- Assign category labels to all project students (or at least to all students for whom reading achievement data is collected).
- If a comparison group is used, assign category labels to all comparison students.
- Assign a category label to each individual student, even if the entire class receives the same reading treatment.
- Record this information in a column adjacent to reading achievement scores.

CATEGORY SYSTEM FOR DESCRIBING READING TREATMENT
IN BILINGUAL PROJECTS
French/English Form

| Category Label | Definition |
|---|---|
| F | Daily reading in French only, for entire year. No English reading. |
| FE | Daily reading in both French and English for entire year. |
| F-FE | Daily reading in French only, from fall to mid-year (sometime between December and March). Daily reading in both French and English from mid-year to end of year. |
| F-E | Daily reading in French only, from fall to mid-year (sometime between December and March). French reading discontinued at mid-year. Transfer to daily reading in English only from mid-year until end of year. |
| E | Daily reading in English only, for entire year. No French reading. |
| E-EF | Daily reading in English only, from fall to mid-year (sometime between December and March). Daily reading in both English and French from mid-year to end of year. |
| ? | Reading treatment unknown. |
| 0 | Other. Please describe. _____ |

Instructions

- Assign category labels to all project students (or at least to all students for whom reading achievement data is collected).
- If a comparison group is used, assign category labels to all comparison students.
- Assign a category label to each individual student, even if the entire class receives the same reading treatment.
- Record this information in a column adjacent to reading achievement scores.

## Format for Reporting Instructional Treatment

When reporting student achievement outcomes, a brief description of the treatment should accompany the data. If results are reported by grade level, then the treatment should also be described by grade level, providing the entire grade level received similar instruction. Report separately achievement outcomes and instructional treatments for those classrooms where an instructional feature or characteristic may have strongly influenced achievement outcomes. For example, if all second grade teachers in a program, except one, were bilingual, report outcomes and describe treatment separately for that one class. Another example: if English reading for a third-grade class differed from all the others because this class was participating in a Title IV reading lab (not part of Title VII program), then this class becomes a separate unit of analysis.

Completed classroom analysis guides (see Section 3-B) are the best source of information for describing instructional treatment. Additional categories of topics to be described can be drawn from the Project Description Model.

The format for describing the treatment could be a chart or a brief narrative. A chart is provided as an example of a format that can be used for summarizing instructional treatment in an evaluation report. If the entire project's instructional treatment were to be described, a separate chart would be used for each subject, for each language group, and for each grade level receiving similar treatment. A completed chart is provided as an example of a description of the instructional treatment for Spanish reading, for limited English speaking students in the second grade.

## Format for Reporting Instructional Treatment

Spanish Reading (Subject) for Students Limited in English Proficiency (Language Group)

Grade Level 2nd grade
Year 1979

No. of students receiving this type of treatment 146
No. of classrooms represented 7

| Subject | Language Use by Teachers and Students | Instructor Characteristics | Subject Taught Hours per Week | Grouping Characteristics | Comments |
|---|---|---|---|---|---|
| Spanish reading. All classes follow project objectives for Spanish reading, and use same basic texts. | Teachers: instruct entirely in Spanish.<br><br>Students: participate almost exclusively in Spanish. English responses are accepted when appropriate. | 6 of the 7 teachers are bilingual. The aide of the 7th class teaches this subject and is bilingual.<br><br>All teachers and aides have received training in teaching Spanish reading. | Taught an average of 6-2/3 hours per week.<br><br>This includes Spanish Language Arts. | All Spanish dominant students receive Spanish reading.<br><br>Reading groups are formed according to achievement level.<br><br>Groups are semi-permanent. | Group N of 146 does not include 16 students who entered the program late in the year. These 16 students were not pretested and will not be included in outcome summaries.<br><br>One classroom was excluded from analysis because bilingual teacher was transferred at mid-year and no bilingual substitute was provided. |

## Format for Reporting Instructional Treatment

_____ for _____

(Subject) (Language Group)

Grade Level _____ No. of students receiving this type of treatment _____

Year _____ No. of classrooms represented _____

| Subject | Language Use by Teachers and Students | Instructor Characteristics | Subject Taught Hours per Week | Grouping Characteristics | Comments |
|---|---|---|---|---|---|
| | Teachers:<br><br>Students: | | | | |

# CHOOSING AN EVALUATION DESIGN

SECTION 4

MAJOR CONTENT ITEM

4-A. A GUIDE TO EVALUATION DESIGNS

## 4. CHOOSING AN EVALUATION DESIGN

Evaluation design is an extremely complex field. Relatively few evaluation specialists have the necessary skills to develop and implement a new, specialized evaluation design. However, choosing from among existing, conventional designs is one of the easiest steps in the evaluation process. This is because there are only a few realistic choices, and, to a large extent, the decisions will be determined by local conditions. The seemingly endless and often esoteric alternatives that provide the subject matter for countless books, articles and conference papers quickly evaporate, in practice. With few exceptions, they are either (a) technically unsound, (b) impossible to implement in a school setting, (c) so complex that only a few experts in the country are qualified to apply them, or (d) some combination of the above.

Two points should be kept in mind in order to avoid unrealistic expectations for evaluation designs. First, even the best designs are not sensitive enough to provide a convincing demonstration of most program impacts. This is simply because program impacts are usually small, and are easily obscured by the effects of many other factors (see Appendix A). However, if large impacts are produced by a program, an appropriate, carefully implemented design will probably provide convincing evidence. Second, most bilingual program evaluation designs are affected by local policies and conditions and by legal and funding agency regulations. In combination, those constraints may completely preclude any accurate assessment of program impact. The only productive option for the project director and evaluator may be to eliminate meaningless impact evaluation activity where regulations and policies permit, and to concentrate on performance-level or other potentially useful information.

Key Problems

Deciding which questions you want to answer. Many evaluations are carried out and reported with no thought as to the exact questions that are being asked or the implications of the answers. Questions that you may wish to ask include: (a) Are students doing better than they would in a conventional classroom? (b) Are they doing better than similar students in other local programs? (c) Are they doing better than similar students in nearby districts, in the entire state, or in the entire country? (d) Are they doing better than last year's program students? (e) Are they doing well enough? (f) Last but not least, has the program improved student performance?

Deciding which questions you can answer. All of the above questions may be of interest, but few evaluations answer all of them equally well. Item 4-A suggests which questions can be answered and suggests some of the problems involved in answering each one.

## Related Issues

Short-term versus long-term evaluations. While most bilingual programs cover several grades, evaluations are often designed as if each grade represented an isolated, complete program. A common example is to test subject matter (e.g., English reading) at grades prior to which the subject has been introduced. While some progress toward the ultimate program goals should be observable at each grade level, the progress may not include all subjects at all grades and the evaluation design should reflect this. Conversely, there is no way to evaluate the total impact of a program until students have completed the whole program. Thus, every bilingual project evaluation should be viewed as a long-term effort.

Overtesting. In addition to the direct costs, testing places a great burden on teachers and students. Simply listing all of the subject areas of interest and finding a test for each one will almost certainly lead to an unreasonable amount of testing. In general, it is advisable to select only the most important areas for special testing and to make use of required district tests wherever possible.

Fall-to-spring versus twelve-month test intervals. A major decision is whether to evaluate impact over a seven-month or a twelve-month period. The shorter period gives a quicker answer and reduces problems of student turnover. However, it also may give a misleading picture due to the short-term impact of some programs, with possible losses over the summer. The twelve-month period is recommended wherever district-policy and other factors permit, since it reduces the testing burden and appears to provide a more meaningful picture of long-term program impacts.

## A Guide to Evaluation Designs
Outline

This section has not been developed beyond the rough outline stage, pending agreement among USOE and potential users as to appropriate content.

### Restricting the Questions

The evaluation design depends on the question:

Question 1. How well are the students performing?
Question 2. How effective is the program compared to previous or alternative programs?

We will not consider:
Do students make achievement gains? (Most students make gains in most subject areas due to maturation.)
Do students meet objectives? (Objectives are arbitrary unless they are based on past years (see "Time series," below) or on other groups of students (see "Comparison groups," below).)

### Answering Question 1 -- Performance Level

This is possible in most districts, at least for subjects with English-language tests.

I. Two kinds of information are required:
   A. Background/experience of the students (See Section 5)
      A general picture of language, schooling, and home/community background is needed if performance levels are to have any meaning at all.
   B. Performance measures (See Sections 2, 6, and 7)
      1. The skills that are measured must be at least generally relevant (to the program goals).
      2. The tests must be reliable, valid, and of the appropriate difficulty.
      3. Testing and scoring must be done with great care.

II. A frame of reference is required. Program students can be compared to any group of interest. Typical comparisons are:
   A. National norms from standardized tests
   B. Local comparison groups
   C. Program students (or, if the program is new -- similar students) from previous years.

III. In addition, <u>interpretation</u> of performance levels requires a description of the relevant instructional treatment.

## Answering Question 2 -- Program Impact

Very difficult or impossible in many districts.

The question is -- What is the effect of the program (separated from other factors)? (a) Is it an improvement over what was being done before? or (b) Is it better than some other program of interest. The "program" here includes procedures materials, and personnel. Separating the effects of personnel from the effects of procedures and materials is an extremely difficult task, and is completely beyond the scope of this manual.

All of the designs listed below assume test data and background/ experience data of the highest quality. Preexisting data from district files are generally not acceptable.

I. Classes of designs using single comparisons.

A. Norm-referenced design

Comparison: National norms from standardized tests.
Measure: Gain in standard score (eg., NCEs) from pretest to posttest.
Problem: Considerable variance among schools (sd = 5 to 10 NCEs) in normal yearly gains or losses.
Credibility: Low.

B. Comparison-group design (non-random assignment)

Comparison: Students in the same district. Both the comparison students and their instruction differ from the bilingual program students. Intended to answer question "b," above.
Measure: Program-student gain compared to comparison-student gain. Pretest differences adjusted by principle-axis adjustment (Section 8). Multiple regression approach is a possible improvement.
Problem: Differences in students, rather than in the program, may produce differences in gains.
Credibility: Low.

C. Time-series design

Comparison: Gains of this year's students are compared to gains of past year's comparable students. Intended to answer question "a," above.
Measure: Program-student gain compared to gain of similar group from preceding years (at least two years are needed). Pretest differences adjusted by principle axis adjustment (Section 8). Multiple regression not needed.
Problem: Changes in school or community may produce differences in gains.
Credibility: Low.

II. Recommended design using a combination of comparisons.

For a credible impact evaluation, all three of the above designs should be combined. Data are required on program-type students plus additional representative groups from the district for several years prior to the start of the new bilingual program. Data on all of the groups should be obtained for several years after the initiation of the new program. This design answers both questions "a" and "b," above, as well as providing information on whether the new program is improving over years.

Norms provide a common metric for comparing groups.

Time series data show changes in the performance of program-type students.

Comparison Groups show that the change is not due to school-wide factors.

Remaining problems:

Shifts in characteristics of program-type students.
Artificial depression of pretest scores, or inflation of posttest scores for program students.

# DESCRIBING STUDENT SKILLS FOR SELECTION AND DATA ANALYSIS

SECTION 5

Major Content Item

5-A. DEMOGRAPHIC AND BIOGRAPHIC INFORMATION WORKSHEET

## 5. DESCRIBING STUDENT SKILLS FOR SELECTION AND DATA ANALYSIS

An accurate description of each student's skills is essential--first, for selecting program students and later, for organizing performance-level and impact data at the analysis stage. Of course, a clear picture of student skills should also guide the instructional program, although instructional planning is beyond the scope of this manual.

The absolute minimum information on each student must include skills in both relevant languages as well as skills in the program's major subject areas. A truly adequate description will also include information on the student's learning background and current environment. The later categories of in formation are especially crucial at the data analysis stage, since they play a major role in determining what can be expected of each student. For example, given a student with a low English reading pretest score, we might expect much greater improvement if the student were a high SES new arrival with no previous training in English reading than we would if the student were from a low SES background and had been in high-quality bilingual programs for several years. Thus, students must be grouped according to both current skills and past experience if meaningful data analyses are to be conducted.

In writing evaluation reports, an accurate description of student background and skills is also essential. Few bilingual-program evaluation reports provide enough information for the reader to make any judgment as to the credibility or importance of the results.

Key Problems

Describing skills in two languages. The problems of measuring language skills is discussed in Section 6. The problem of concern here is simply that many evaluations ignore target-language skills. Selection, instruction, and data analysis are often based only on the fact that students have limited English skills. In some projects, the implicit assumption of superior skills in the target language is entirely justified. In many of the projects observed by RMC, however, target language skills were even lower than English skills, sometimes substantially so. If such situations are not made clear in the evaluation report, the results become completely misleading.

Describing student backgrounds. A clear picture of the program students' environment and learning history are also essential to the accurate understanding of impact evaluation results. Apparently few projects collect this information for impact evaluation purposes, and even fewer present a systematic treatment of this information in their evaluation reports. While this manual does not provide a complete guide to the appropriate use of such information, the Biographic and Demographic Worksheet in this section should provide a starting point. (See also: Data Analysis, Section 8.)

## Related Issues

Combining measures into a single, selection score. Selection for a bilingual program may be based on student background, categories, language test scores, achievement test scores, and teacher judgment. If the school district is willing to quantify all of these measures (including teacher judgment), arrive at a single score, and use this score as the sole basis for assigning students to the program, then statistical corrections to achievement gains become possible, and the accuracy of the achievement impact evaluation may be considerably improved.

Longitudinal student profiles. Since most bilingual programs span several grade levels, the value of student descriptions is increased greatly by creating longitudinal student profiles. Since most schools keep permanent student record files, it may only require minor additions to ensure that the appropriate background (and treatment) information is readily available for each program student.

Selecting and describing students who are proficient in English. Many programs include substantial numbers of monolingual, native-English speakers and bilinguals who are highly proficient in English. For these students, it is not necessary to maintain the same amount of information on English-language experience. Of course, these students must be analyzed separately from those who are learning English as a second language.

## Demographic and Biographic Information Worksheet

This demographic and biographic information worksheet can be used to document information which will enable the evaluator to interpret results with a higher degree of accuracy. The information gathered can also be used for individual pupil records so as to facilitate continuity of instruction across the years.

The demographic information should be collected per school, whereas the biographic information must be collected for each student. Similar information should be collected for control or comparison students, if the evaluation design incorporates a control or comparison group.

School Year _____

School District _____

School(s) _____

_____

_____

_____

Information Collected by _____

Date _____

Demographic and Biographic Information Worksheet
(Check appropriate answer in margin)

I. School/Community Characteristics

A. These school/community characteristics apply to:
(Use one form per group.) A.
1) treatment students 1)_____
2) comparison students 2)_____

B. How is project student defined?

_____
_____
_____

C. What percentage of the students in the project school(s) are in the Title VII project?
(List by school name or code; put percentage in margin.)

| School name or code | C. |
|---|---|
| 1) _____ | 1) _____ |
| 2) _____ | 2) _____ |
| 3) _____ | 3) _____ |
| 4) _____ | 4) _____ |
| 5) _____ | 5) _____ |

D. Do(es) the school(s) participate in the free lunch program? (write yes or no in margin)

| School | D. |
|---|---|
| 1) _____ | 1) _____ |
| 2) _____ | 2) _____ |
| 3) _____ | 3) _____ |
| 4) _____ | 4) _____ |
| 5) _____ | 5) _____ |

E. What criteria are used to determine a student's English language proficiency classification (e.g., LEP/LESA)? (Check appropriate answer in margin.) E.

1) Teacher judgment 1) _____
2) Language proficiency test 2) _____
Test:_____
Cutoff:_____
3) Achievement test 3) _____
Test:_____
Cutoff:_____
4) Combination of the above (specify) _____ 4) _____
5) Other (explain) _____ 5) _____

F. At the time of fall testing, what percentage of the students in the project were classified as limited in English proficiency? F. _____

G. What is the size of the district in which the project is located? (Check appropriate answer in margin.)

| | G. |
|---|---|
| 1) 12,000 or more students | 1) _____ |
| 2) 3,000 to 11,999 | 2) _____ |
| 3) 1,000 to 2,999 | 3) _____ |
| 4) less than 1000 | 4) _____ |

H. In what type of community is the project located? (Check appropriate answer in margin.)

| | H. |
|---|---|
| 1) Metropolitan | 1) _____ |
| 2) Urban | 2) _____ |
| 3) Suburban | 3) _____ |
| 4) Rural | 4) _____ |

I. What percentage of the community is Hispanic (or of the ethnic group being served by the program)? I. _____

II. Pupil Characteristics (Use one form per student)

A. For record-keeping purposes, what is this pupil's: A.
1) Code number 1) _____
2) Age (as of fall of 19__) 2) _____
3) Ethnicity 3) _____
4) Language classification: 4) _____
a) Limited in English proficiency
b) Proficient in English

B. Which group does this pupil belong to?
(Check appropriate answer in margin.) B.
1) Treatment 1) _____
2) Comparison 2) _____

C. In what country was the student born?
(Check appropriate answer in margin.) C.
1) United States 1) _____
2) Spanish speaking country (or country where target language is spoken) 2) _____
3) Other (Please specify) _____ 3) _____
4) Unknown 4) _____

D. To the best of your knowledge how long has the student been in the U.S., as of fall 19__ (current year)? D.
1) Less than one year 1) _____
2) One to two years 2) _____
3) More than two years 3) _____

E. What is the number of years of schooling this student has completed outside the U.S.? E.
1) Years of schooling outside U.S. _____ 1) _____
2) Don't know _____ 2) _____
3) N/A _____ 3) _____

F. What language is most frequently spoken to the student at home? F.
1) Spanish (or non-English program language) 1) _____
2) English 2) _____
3) Equal use of two languages 3) _____
4) Other _____ 4) _____

G. What language does the student use most frequently at home? G.
1) Spanish (or non-English program language) 1) _____
2) English 2) _____
3) Equal use of two languages 3) _____
4) Other _____ 4) _____

H. How was the information in Item E. obtained? H.
1) Parent survey 1) _____
2) Self report (child) 2) _____
3) Teacher/staff judgment 3) _____
4) Other _____ 4) _____

I. How was the information in Item F. obtained? I.
1) Parent survey 1) ____
2) Self report (child) 2) ____
3) Teacher/staff judgment 3) ____
4) Other _____ 4) ____

J. In which of the following programs is the student currently participating? (Please check.) J.
1) Free lunch program 1) ____
2) Title I 2) ____
3) Migrant 3) ____
4) ESAA 4) ____
5) Other (Please specify.)
_____ 5) ____

K. Indicate the number of years this student has participated in bilingual education programs, prior to current year. K. ____

L. How would you characterize this student's absentee rate? L.
1) seldom = approximately __ day(s)/month 1) ____
2) average = approximately __ day(s)/month 2) ____
3) frequent = approximately __ day(s)/month 3) ____

M. What percentage of the children in the student's classroom of instruction (at the time of fall testing) were considered students of limited English proficiency? M. ____

# SELECTING TESTS

SECTION 6

Major Content Items

6-A. SELECTING AN ACHIEVEMENT TEST

6-B. SELECTING A LANGUAGE PROFICIENCY TEST

6-C. A SYSTEM FOR COMPARING CURRICULUM CONTENT WITH THE CONTENT OF CTBS SPANISH AND ENGLISH, FORMS B AND C

## 6. SELECTING TESTS

Although catalogues of tests list thousands of titles, the selection of tests for a bilingual-program evaluation actually involves few real choices. This is because (a) federal, state, and local regulations largely determine the subject areas to be tested and often the pool of acceptable tests as well, and (b) only a handful of the available tests meet minimum technical requirements. While satisfactory tests are available for basic subject areas, the perfect test does not exist, and searching for such tests among the more obscure titles is an expensive exercise in futility.

The major concern in selecting tests is to be sure that all major program goals are covered (i.e., all major subjects, at all relevant grades, and, where warranted, in both languages). Tests must meet reasonable standards of reliability and validity. It is also advisable to check the technical manual to see that the test publisher has employed procedures designed to reduce culture and linguistic bias.

Key Problems

Matching the tests to the program. The minimum matching requirements are simply to test the subjects that are included in the bilingual program and not to test specific subject matter before it has been introduced. Following these two simple and obvious rules would drastically improve many evaluations. A more thorough matching process is advisable, and is addressed in Item 6-C.

Language tests for selection, diagnosis, and impact evaluation. The best ways to design selection and diagnostic tests are still highly controversial subjects among language test developers, and there are problems with all such tests that are currently available. Even greater problems arise when using selection or diagnostic tests to measure language improvement. These problems are noted in Item 6-B.

Other Issues

Selecting tests with non-English language norms. Basically, non-English language norms adequate for impact evaluation do not exist. The Inter-American Tests provide user-norms based on students in bilingual programs using that test. The norms provided with the Spanish CTBS do not represent the population of Spanish/English bilingual students. Norms for both tests can only provide a possible standard for performance-level evaluations. (See Item 6-A.)

Test level (floor and ceiling effects). In some bilingual programs, the at-grade-level test is too difficult for program students at pretest The next lower level may be too easy at posttest time. If the mean score on a test is less than 25 percent of the items correct or more than 75 percent of the items correct, floor or ceiling effects probably exist. See Item 6-A.

Longitudinal and multi-grade-level requirements. Most bilingual programs offer several grade levels. Therefore, it is desirable to have achievement tests that can be compared across grades and that can be used to follow groups of students as they progress through the grades. In practice, this means using one of the well known achievement tests from the major publishing companies. See Item 6-A.

Criterion-referenced tests (CRTs). In recent years, CRTs have been advocated widely as a solution to the many problems of standardized (norm-referenced) tests. In fact, the advocates of CRTs have not solved the problems. They have merely attempted to avoid them by asking different kinds of evaluation questions. Where the basic question concerns program impact, the reliability and validity of the tests are of primary importance. Rephrasing the impact questions in CRT terminology simply helps to obscure or ignore the fundamental reliability and validity problems. Although, in principle, CRTs can be just as reliable and valid as norm-referenced tests (in fact, a single test can be both norm and criterion referenced) in practice, CRTs often lack reliability and are likely to reduce the accuracy of an impact evlauation.

Language of testing. There are no definitive guidelines as to which language should be used for testing subjects other than language (i.e., math, science, culture). If students are very weak in one language, it seems obvious that that language is inappropriate for testing. Some PIP field-test sites in which students were reasonably skilled in both languages tested math in both languages. In these sites, the language of testing had little effect on scores. In general, the language of testing should be determined after considering the goals of the program and the language of instruction, as well as the language proficiency of the students.

Must English and non-English language tests come from the same publisher? This question applies mainly to Spanish-English programs, since few tests are available in other languages. While there are some advantages to dealing with a single test publisher, it is more important to get the most appropriate tests in each language. Limiting choices to tests that are published in two languages is an unnecessary restrictions.

## Selecting an Achievement Test

In selecting achievement tests for the evaluation of bilingual programs, evaluators must consider all the same criteria that are used in selecting any achievement test as well as additional criteria that relate to the nature of the program and the student population. This discussion will give most emphasis to issues in test selection that are especially important for bilingual education evaluations.

### Test Bias

During the last ten years extensive attention has been given to the effects of test bias for culturally different populations (Wargo, 1975; Houts, 1977). As a result, test publishers have made concerted efforts in this area and many standardized achievement tests have been revised. The technical manual of a test will often include a discussion of what procedures were undertaken to minimize bias. The two most common procedures are: (1) review of the content of the items by a culturally sensitive panel and (2) statistical item analysis.

Review of content. Reading and examining the content of items may result in rewriting items so that they seem fairer to all groups involved. However a visual examination alone cannot determine if an item is biased, i.e., that it will function differently for different groups of students. What can be accomplished is the elimination of stereotypcial wording or content. External review panels have the advantage of insuring a disinterested reading, although in-house groups may also be effective. This procedure may result in a more acceptable test, but will not necessarily eliminate biased items.

Item analysis. Item analysis is a statistical procedure that is performed routinely in test construction. The scores of students on each item are compared to their scores on the whole test in order to determine if each item is measuring what the whole test measures, and in fact should be part of that test. When this procedure is used to eliminate bias towards

a specific group, the test is administered to both the general population and to the specific group. Then item analysis is performed in order to determine that the same items function similarly for both groups. For example, if item 1 is difficult for one group it should be difficult for the other regardless of the mean test scores for each group. If an item is easy for one group but difficult for another, then such an item exhibits bias, and should probably be eliminated.

Additional Selection Issues

Consideration of subtest content and weight in scoring is important for selecting the test that most closely matches the curriculum and for determining whether in-level testing is appropriate. Such issues are important for all students, but they may be even more critical for students of limited English proficiency. Although the curriculum of bilingual programs may contain the same final objectives, skills such as English reading may not be taught in the same grade levels as other programs.

The wording of the instructions to the test should be considered. The language of the instructions should not be more difficult than the language used in the items that actually appear in the test. Although directions containing needlessly complex sentence structure are a handicap for all students, they will cause an even greater difficulty for students of limited English proficiency. Examiners may want to consider systematically simplifying test directions, but if norms are to be used, this may affect their validity.

Additionally, the content of the test should be examined to determine the extent to which it tests the out-of-school experience of the children. The experience of the culturally different child and of the low SES child may differ significantly from that assumed by the authors of the test. Therefore, the more the test relies on out-of-school experience, the more it may discriminate against the target population and the less valid it will be for evaluating program impact.

Finally, if bilingual tests are used, the nature of the translation should be considered. Some tests are direct translations except where such a translation would clearly be impossible. Other tests provide equivalent versions where the kinds of items and the difficulty level are roughly equivalent, but the content of the item may be completely different. Other tests are a combination of both methods. In a translated test, the difficulty level may not be the same for both versions. However, very few test publishers provide equivalent versions.

Language of Testing

In many bilingual education evaluations, the evaluator must decide in what language to test. Several questions have to be considered individually and in relation to each other. First, what is the language of instruction for the subject that will be tested? Because the language of instruction for math, for example, may be different for students in the same class or may be different at various times during the year, this question may not be answered simply. Second, what is the dominant language of the child as established by a systematic assessment procedure? Third, what are the project goals? Goals may require testing in a particular language. Ideally, of course, students should be tested in the language in which they will do best. However, that language may not always be the dominant one. For example, a student may be more fluent in Spanish, but if almost all math instruction has been in English, the student may perform better on an English test.

There are other issues involved in planning testing in more than one language that have not yet been studied in sufficient detail. Some evaluators double test the project students, avoiding the choice of test language by testing in both languages. The benefits of this practice are clear: more information is obtained about the students' proficiency in content and language and the dangers of testing only in the weaker language are avoided. However, the additional expense, the added burden on teachers and students, and the possibility of practice effects represent significant disadvantages. In addition, the language of some students may be neither standard English nor standard Spanish.

Where tests exist in two languages, Spanish may be the most appropriate language for the pretest; but, after a year of English instruction, English may be the most appropriate language for the posttest. Longitudinal studies will almost certainly include scores in both languages reported at different stages of a student's progress. Evaluators will have tc consider carefully the interpretations of such scores.

## Limits to the Usefulness of Norms

The use of national norms as a comparison standard in an evaluation relies on the validity of a principle known as the equipercentile assumption. This assumption implies that in the absence of any special instructional treatment students in the project would have grown at a rate comparable to that of students in the norming sample who obtained the same mean pretest value. Such an assumption can only be valid if the project population is similar in educationally relevant ways to the population represented in the norming sample. This is not usually the case in bilingual education programs which are generally comprised of students of limited English proficiency, bilingual students, and a larger proportion of low SES students than is found in the general population. While the accuracy of the equipercentile assumption for such populations has not yet been systematically assessed, it is unlikely that norms for English achievement tests can provide precise no-treatment expectations for bilingual project students. There are no statistical techniques to adjust for differences in expected growth between the project students and the norming population (Tallmadge, 1976).

Recently data have been gathered on Spanish language achievement tests. The most recent editions of the Comprehensive Tests of Basic Skills (CTBS) and the Inter-American Series both furnish norms tables for English and Spanish versions of their tests, but the manner in which such norming data were compiled limits their usefulness for evaluating the impact of bilingual projects. The CTBS Espanol norms were developed by administering the CTBS in both languages to a balanced bilingual, biliterate population as determined by scores on the SERVS test. The assumption was made

that a student's standing in the norms would be the same in English and Spanish. Students' scores in Spanish were then equated with their rank in the English norms. Although the assumption that a perfectly bilingual person will possess the same knowledge of content in two languages is logical, the possibilities for error are so large that the Spanish norm conversions can provide only very rough estimates of student achievement. There are several other reasons why the CTBS norms cannot be used to provide a precise estimate of project impact. Because the scores in the norms table are extrapolated rather than derived empirically, they are subject to a certain amount error inherent in any estimation procedure. In addition, the balanced bilingual population in the sample is not comparable to the population of most bilingual programs which include students with a range of language proficiencies. Finally, because the students in the sample were in bilingual programs they do not provide an estimate of how similar students would have performed without any special instruction.

The Inter-American norms were not constructed from a national probability sample. They are "user norms" derived only from those groups in the population to whom the Inter-American tests were administered in the course of local evaluations. For certain tests, the sample obtained in this way numbers over a thousand students, but for others the N is less than 100, severely limiting the reliability of normative data, particularly in the extreme score ranges where estimates are based on relatively few cases. Because the norming group was not specifically constructed to represent the population of limited English and bilingual students, unknown biases may exist in the sample. Because students in the sample are also in bilingual programs, the norms do not provide an estimate of how similar students would have performed in the absence of a special program.

The question of how a group of students would have performed without a bilingual project cannot be answered by simply consulting currently available norms. But existing norms can be used to answer other evaluation questions. Well constructed norms based on national probability samples, such as those provided by the major achievement tests, can be used to show how the bilingual project students compare to national averages. Norms based on more specific populations, such as those constructed

for the Spanish versions of the CTBS and the Inter-American, can be used to show how project students compare to the bilingual/bil'terate CTBS sample or the bilingual project students in the Inter-American sample.

Out-of-level testing. The use of tests at levels below those recommended by the publisher is an option if the content of the program can be measured better this way. Students in bilingual programs may be learning skills, such as English reading, at a later time than other students and therefore, should receive the same test at a later point. In order for any test to be suitable, the average score of the group tested should be between 1/3 and 3/4 of the maximum (Roberts, 1976). Otherwise, ceiling or floor effects depress estimates of student gains. Some publishers provide norms for the administration of a single test in several grades. Other publishers provide expanded standard scores that link up all levels of a test on a common scale, and occassionally, locator tests, to facilitate out-of-level testing. Generally, a test should be used no more than one level below that recommended by the publisher. But care should be taken that in testing out-of-level, pretest floor effects are not being replaced by posttest ceiling effects.

Introduction to test list. An extraordinary number of tests could be used to evaluate basic subject areas for bilingual programs. Some of these tests are locally developed and have not been administered to large samples of the population. Therefore, they are less likely to have the technical qualities required by most evaluators. Other tests are limited to only one content area, and cannot be used by themselves to evaluate a bilingual project which includes several content areas. Finally, many evaluators will first consider the appropriateness of tests already in use in the district for the evaluation of the bilingual program. Certain tests may be mandated or choices may be constrained in other ways. Selection of a test already being used for district-wide assessment intorduces the possibility of comparison with local non-project students. This comparison alone cannot provide a precise estimate of project impact, but may answer other evaluation questions, such as how project students compare in achievement level and rate of growth to other students in the district.

The annotated test list which follows is an attempt to provide helpful information about tests that, for the reasons discussed above, are already likely to be under consideration by project evaluators.

Only major tests of achievement that include both math and reading or language subtests were considered. All such tests available in two languages were included. Tests only available in English were limited to those included in the Anchor Test Study (Loret, 1974). Finally, all of the tests were discussed only as they apply to evaluations of grades K-6.

The same categories of information are provided for each test to facilitate comparison. All of the tests are available from major publishers. Technical aspects of such tests are likely to be as good as the state-of-the-art. All of the tests have technical manuals describing the process of test construction and standardization. Except for an occasional subtest, all of the tests are designed to be administered in groups. Administration time for each test varies according to the number of subtests used. Subtests are listed only where they contribute to a total score in reading, languge arts, or mathematics, three major areas of interest to bilingual program evaluation.

## REFERENCES


Hoepfner, R. Achievement test selection for program evaluation. In In Wargo, M. J. and Green, O. R. (ed.), Achievement testing of disadvantaged and minority students for educational program evaluation. CTB/McGraw Hill, 1977.

Houts, P. L. The myth of measurability. New York: Hart Publishing Company, Inc., 1977.

Loret, P. G., et al. Anchor test study. Washington: U.S. Government Printing Office, 1974.

Rhodes-Hoover, M., Politzer, R. L., & Taylor, O. Bias in achievement and diagnostic reading tests: A linguistically oriented view. Unpublished manuscript, Stanford University, 1975.

Roberts, A. O. H. Out-of-level testing. Mountain View, CA: RMC Research Corporation, 1978.

Tallmadge, G. K. Cautions to evaluators. In Wargo, M. J. and Green, O. R. (ed.), Achievement testing of disadvantaged and minority students for educational program evaluation. CTB/McGraw Hill, 1977.

Wargo, M. J., & Green, D. R. Achievement testing of disadvantaged and minority students for program evaluation. CTB/McGraw Hill, 1977.

California Achievement Test, 1977-78
Forms C and D

1. Languages: English

2. Publisher's recommended in-level use:

| Level | Grade |
|---|---|
| Level 10 | K.0-K.9 |
| Level 11 | K.6-1.9 |
| Level 12 | 1.6-2.9 |
| Level 13 | 2.6-3.9 |
| Level 14 | 3.5-4.9 |
| Level 15 | 4.5-5.9 |
| Level 16 | 5.5-6.9 |

3. Subtest Components:

| Level: | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|
| Pre-reading | | | | | | | |
| Listening for Information | X | | | | | | |
| Letter Forms | X | | | | | | |
| Letter Names | X | | | | | | |
| Letter Sounds | X | | | | | | |
| Visual Discrimination | X | | | | | | |
| Sound Matching | X | | | | | | |
| Reading | | | | | | | |
| Vocabulary | | X | X | X | X | X | X |
| Comprehension | | X | X | X | X | X | X |
| Phonic Analysis | | X | X | X | | | |
| Structural Analysis | | | X | X | | | |
| Language Total | | | | | | | |
| Language Mechanics | | X | X | X | X | X | |
| Language Expression | | X | X | X | X | X | |
| Mathematics Total | | | | | | | |
| Computation | | X | X | X | X | X | X |
| Concepts and Applications | | X | X | X | X | X | X |

4. Norming: Weeks rather than midpoint dates are provided for empirical fall and spring norms. These are the week in which November 3rd falls, and the week in which May 4th falls. Tests can be administered two weeks on either side of these weeks without the use of interpolated norms.

5. Out-of-level testing: Provides an expanded standard score scale and a locator test.

6. Procedures for minimizing bias: Test writers followed guidelines to avoid bias in the development and editing of items. Items were reviewed by representatives of various ethnic and cultural groups. An extensive item analysis was conducted with the tryout items to compare responses of "Black" students and "other" students. A point biserial correlation was used to show the relation of items to category objective scores, and grade-to-grade growth as shown by item difficulties was also examined. The percent of biased items found in the trial items for the various subject areas ranged from 25 to 7 percent. After revision the percent of biased items was reduced to the 3-0 percent range.

CIRCUS
1976

1. Languages: English

2. Publisher's recommended in-level use:

| Level | Grade |
|---|---|
| Circus A | Nursery School and Kindergarten - Fall |
| Circus B | Kindergarten - Spring |
| | First Grade - Fall |
| Circus C | First Grade - Spring |
| | Second Grade - Fall |
| Circus D | Second Grade - Spring |
| | Third Grade - Fall |

3. Subtests:*

| | Level | | | |
|---|---|---|---|---|
| | A | B | C | D |
| Pre-reading | | X | | |
| Reading | | | X | X |
| Listen to the Story | X | X | | |
| Listening | | | X | X |
| How Much and How Many | X | X | | |
| Mathematics | | | X | X |
| Writing Skills | | | | X |

*Many other subtests are provided, but only these that coordinate with the STEP are listed here. No total scores are possible from any combination of subtests.

The subtests listed above provide coordination through content and expanded standard scores with the following subtests of STEP III, Level E-J; Reading, Listening, Math Concepts and Math Computation, and Writing Skills.

4. Norming: The Circus was administered to a national probability sample during the fall (October) only. Therefore, the comparison of a group to the national sample for pre- and posttesting can be done for a fall-to-fall evaluation design only. Information is also provided in sentence form describing what each range of scores means in terms of skills mastered. A fall to spring comparison of the proportion of students falling in each category could be made, but would require the use of a local comparison group to determine the normal-growth expectation. Separate tables exist for comparing groups and for comparing individuals. The normative data is very well suited to individual student evaluation because the national sample is divided into subgroups such as sex, geographic region, and SES.

5. Out-of-level testing: Expanded standard scores can be used for subtests that coordinate with STEP III.

6. Procedures for minimizing bias: No statistical procedures are reported. Separate norms are provided according to categories such as sex, geographic region, and SES.

EL CIRCO
1979

1. Languages: Spanish and English

   Spanish tests allow the test administrator to select among alternatives the word most appropriate for the students' variety of Spanish.

2. Publisher's recommended in-level use: Tests can be used at preschool, kindergarten, and beginning of first grade.

3. Subtests:*

   Cuanto y Cuantos

   Para Qué Sirven Las Palabras

   What Words are For

   Quanto y Quantos is a direct translation of Level A of How Much and How Many of CIRCUS. Para Que Sirven Las Palabras and What Words are For are equivalent, but one is not a translation of the other. For example, each test has items testing comprehension of the past tense but the items will have a different content.

4. Norming: The El Circo measures were administered to a nationwide sample of children from the Spanish-speaking cultural groups. Empirical norms exist for fall only.

5. Out-of-level testing: Separate norms exist for preschool, kindergarten, and first grade.

6. Procedures for minimizing bias: Items were reviewed by a cultural advisory committee composed of speakers of Puerto Rican, Mexican, and Cuban Spanish.

---

*Several tests have been developed as part of El Circo, but only the ones listed are available for spring 1980.

Comprehensive Test of Basic Skills
English Version 1973, Spanish Version 1978
Form S

1. Languages: English and Spanish
The CTBS/Español is a direct translation of the English CTBS/S with the exception of certain items which could not be translated or which required different translations for dialects of Spanish. In such cases equivalent items have been constructed.

2. Publisher's recommended in level use:

| | English CTBS/S | CTBS/Español |
|---|---|---|
| Level B | Grades K.6-1.9 | Grade 1 |
| Level C | Grades 1.6-2.9 | Grade 2 |
| Level 1 | Grades 2.5-4.9 | Grades 3 and 4 |
| Level 2 | Grades 4.5-6.9 | Grades 5 and 6 |

3. Subtest components:

| | Level | | | |
|---|---|---|---|---|
| Component | B | C | 1 | 2 |
| Reading | | | | |
| Word Recognition | X | | | |
| Reading Vocabulary | | X | X | X |
| Reading Comprehension | X | X | X | X |
| Mathematics | | | | |
| Math Computations | X | X | X | X |
| Concepts & Applications | X | X | X | X |

4. Norming: The norms for the Spanish version of the CTBS were derived through a spring equating with the nationally representative English language norms. The no-treatment expectation obtained by their use is not referenced to a Limited English Proficiency population but rather to the English language performance that could be expected from the bilingual/biliterate population on whom the equating was done. The scoring patterns in both English and Spanish for limited English proficiency students may be quite different; therefore, the norms do not present a precise standard of comparison. Empirical norms exist for the English CTBS for spring for grades 2-6, and for fall and spring for grades K and 1.

5. Out-of-level testing: An expanded standard score scale is available for the CTBS/S norms.

6. Procedures for minimizing bias: Prior to standardization items were reviewed by Black and Spanish-speaking consultants. In addition, trial items were administered to a sample of Black students and "other" students. Items with a point-biserial coefficient of less than .2 were rejected. A subsequent analysis was made of the test results of Black students, Spanish-speaking students, and other students. Although the mean scores were lower for the Black and Spanish-speaking group, the tests appeared to be functioning similarly for both groups.

Inter-American Series: Test of Reading, 1962-69
Forms CE, DE, CEs, DEs

1. Languages: English, Spanish, and French
   Spanish version is an exact translation of English version.

2. Publisher's recommended in level use:

| | |
|---|---|
| Level 1 | Grade 1.5-2.5 |
| Level 2 | Grade 2.5-3.9 |
| Level 3 | Grades 4,5,6 |

3. Subtest components:

| | Level | | |
|---|---|---|---|
| Components | 1 | 2 | 3 |
| Vocabulary | X | X | X |
| Comprehension | X | | |
| Level of Comprehension | | X | X |
| Speed of Comprehension | | X | X |

4. Norming: The Inter-American norms were not developed using a probability sample. They are based on data collected from test users. The test manual states that these norms "should be applied with caution until local norms can be developed." Although N's for some tests consist of more than a thousand students, others comprise less than a hundred students. For these reasons, the norms do not provide a convincing, precise standard of comparison.

5. Out-of-level testing: Norms are provided for out-of-level testing; however, above comments regarding norms should be taken into account.

6. Procedures for minimizing bias: Content was selected that is familiar to English and Spanish speakers of the Western Hemisphere. A semantic frequency list was consulted in wording the translation, but the manual states that frequency is not always an indication of difficulty level. Spanish trial items were administered to Spanish speakers, and English trial items were administered to English speakers, after which item analysis and item selection were performed.

## Inter-American Series: Test of General Ability, 1961-72
## Forms CE, DE, CEs, and DEs

1. Languages: English and Spanish
   Spanish version is an exact translation of English version.

2. Publisher's recommended in-level use:

| | |
|---|---|
| Preschool Level | Ages 4 and 5 |
| Level 1 | Grades end K, Grade 1 |
| Level 2 | Grades 2, 3 |
| Level 3 | Grades 4, 5, 6 |

3. Subtest components:

| | Level | | | |
|---|---|---|---|---|
| Components | Pre-School | 1 | 2 | 3 |
| Oral Vocabulary | X | X | X | |
| Number | X | X | X | |
| Association | X | X | | |
| Classification | X | X | X | X |
| Analogies | | | X | X |
| Sentence completion | | | | X |
| Computation | | | | X |
| Word Relations | | | | X |
| Number Series | | | | X |

4. Norming: The Inter-American norms were not developed using a probability sample; the norms are based on data collected from test users. The test manual states that these norms "should be applied with caution until local norms can be developed." Although N's for some tests consist of more than a thousand students, others comprise less than a hundred students. For these reasons, the norms do not provide a convincing, precise standard of comparison.

5. Out-of-level testing: Norms are provided for out-of-level testing; however, above comments regarding norms should be taken into account.

6. Procedures for minimizing bias: Content was selected that is familiar to English and Spanish speakers of the Western Hemisphere. A semantic frequency list was consulted in wording the translation, but the manual states that frequency is not always an indication of difficulty level. Spanish trial items were administered to Spanish speakers, and English trial items were administered to English speakers, after which item analysis and item selection were performed.

IOWA Tests of Basic Skills, 1978
Forms 7 and 8

1. Languages: English

2. Publisher's recommended in-level use:

| Level | Grade | Forms |
|---|---|---|
| Primary Battery 5 | K.1-1.5 | 7 |
| Primary Battery 6 | K.8-1.9 | 7 |
| Primary Battery 7 | 1.7-2.6 | 7 |
| Primary Battery 8 | 2.7-3.5 | 7 |
| Multilevel Battery 9 | 3 | 7 and 8 |
| Multilevel Battery 10 | 4 | 7 and 8 |
| Multilevel Battery 11 | 5 | 7 and 8 |
| Multilevel Battery 12 | 6 | 7 and 8 |

3. Subtest components:

| | Level | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Reading | | | | | | | | |
| Reading Comprehension | | | | | X | X | X | X |
| Pictures | | | X | X | | | | |
| Sentences | | | X | X | | | | |
| Stories | | | X | X | | | | |
| Reading | | X | | | | | | |
| Vocabulary | X | X | X | X | X | X | X | X |
| Math | | | | | | | | |
| Math Concepts | | | X | X | | | | |
| Math Problems | | | X | X | | | | |
| Math Computations | | | X | X | X | X | X | X |
| Math | X | X | | | | | | |
| Language | | | | | | | | |
| Spelling | | | X | X | X | X | X | X |
| Capitalization | | | X | X | X | X | X | X |
| Punctuation | | | X | X | X | X | X | X |
| Usage | | | X | X | X | X | X | X |
| Language | X | X | | | | | | |
| Listening | X | X | X | X | | | | |

4. Norming: Empirical norms exist for 15 October and 15 April.

5. Out-of-level testing: An expanded standard score scale is provided.

6. Procedures for minimizing bias: Authors with diverse cultural backgrounds participated in writing of test.

Metropolitan Achievement Tests
(MAT) 1978 Forms J1 and K1

1. Languages: English

2. Publisher's recommended in-level use:

| Level | Primary |
|---|---|
| Primer | K.5-1.4 |
| Primary 1 | 1.5-2.4 |
| Primary 2 | 2.5-3.4 |
| Elementary | 3.5-4.9 |
| Intermediate | 5.0-6.9 |

3. Subtest components

| | Primer | Primary 1 | Primary 2 | Elemen-tary | Interme-diate |
|---|---|---|---|---|---|
| Reading Comprehension* | X | X | X | X | X |
| Language | | | | | |
| Listening Comprehension | X | X | X | X | |
| Punctuation and Capitalization | | X | X | X | X |
| Usage | | X | X | X | X |
| Grammar and Syntax | | X | X | X | X |
| Spelling | X | X | X | X | X |
| Study Skills | X | X | X | X | X |
| Math | | | | | |
| Numeration | X | X | X | X | X |
| Geometry and Measurement | X | X | X | X | X |
| Problem Solving | | X | X | X | X |
| Operations: Whole Numbers | X | X | X | X | X |
| Operations: Laws and Properties | | | | X | X |
| Operations: Fractions & Decimals | | | | | X |
| Graphs & Statistics | | | | | X |

*Additional reading subtests such as rate and auditory discrimination are available, but they are not part of the comprehension score.

4. Norming: Empirical fall and spring norms have been developed with midpoints of 15 October and 20 April respectively.

5. Out-of-level testing: Provides an expanded standard score scale. Out-of-level testing should be no more than one level below that recommended for the grade.

6. A combination of objective and subjective methods was used to identify ethnically biased items on the MAT. Following review by a panel of ethnically diverse educators, test items were examined for bias using three conceptually different statistical detection methods. Items tagged as biased by either the subjective or objective procedures were subsequently revised or eliminated.

## Sequential Tests of Educational Progress (STEP) III, 1979, Forms X and Y

1. Languages: English

2. Publisher's recommended in level use:

| Level | Grade |
|---|---|
| Intermediate E | 3.5-4.5 |
| Intermediate F | 4.5-5.5 |
| Intermediate G | 5.5-6.5 |

3. Subtest components:

| | Level | | |
|---|---|---|---|
| | E | F | G |
| Reading Total | | | |
| Vocabulary | X | X | X |
| Comprehension | X | X | X |
| Inference | X | X | X |
| Math | | | |
| Mathematics Basic Concepts | X | X | X |
| Mathematics Computations | X | X | X |
| Language: Writing Skills | | | |
| Spelling | X | X | X |
| Capitalization | X | X | X |
| Word Structure and Usage | X | X | X |
| Sentence and Paragraph Organization | X | X | X |
| Language: Listening | | | |
| Listening Comprehension | X | X | X |
| Following Directions | X | X | X |

4. Norming: Empirical norms are available for fall and spring. Midpoints of the norming periods are 5 October and 10 May.

5. Out-of-level testing: Provides expanded standard score scale and also out-of-level norms. Has locator test.

6. Procedures for minimizing bias: Items were edited by in-house minority and women test specialists, and by an external minority review panel.

7. Additional comments: Can be used in conjunction with CIRCUS, 1978, because of the coordination of test content and an expanded standard score scale.

SRA Achievement Series, 1978, Forms 1 and 2

1. Languages: English

2. Publisher's recommended in level use:

| Level | Primary |
|---|---|
| A | K.5-1.5 |
| B | 1.5-2.5 |
| C | 2.5-3.5 |
| D | 3.5-4.5 |
| E | 4.5-6.5 |

3. Subtest components:

| | Level | | | | |
|---|---|---|---|---|---|
| Component | A | B | C | D | E |
| Reading | | | | | |
| Visual Discrimination | X | | | | |
| Auditory Discrimination | X | X | | | |
| Letters/Sounds | X | X | X | | |
| Listening Comprehension | X | X | X | | |
| Vocabulary | | X | X | X | X |
| Comprehension | | X | X | X | X |
| Mathematics | | | | | |
| Concepts | X | X | X | X | X |
| Computation | | X | X | X | X |
| Problem Solving | | | | | X |
| Language Arts | | | | | |
| Mechanics | | | X | X | X |
| Usage | | | X | X | X |
| Spelling | | | X | X | X |

4. Norming: The norms are based on a nationally representative sample of students. Empirical spring norms are available with temporary fall interpolated norms. Empirical fall norms are currently being developed. Empirical fall and spring norming dates are: 7 October and 25 April.

5. Out-of-level testing: Out-of-level testing can be interpreted using the SRA expanded standard score scale known as GSV (Growth Scale Value).

6. Procedures for minimizing bias: Items were edited by representatives of minority groups and women. The trial items were administered to a sample that included Black, Hispanic, American Indian and non-minority subsamples. The items were then examined statistically and items which were easy for one group, but difficult for another were eliminated.

Stanford Achievement Test, 1973
Forms A, B, and C

1. Languages: English

2. Publisher's recommended in level use:

| Level | Primary |
|---|---|
| Primary I | 1.5-2.4 |
| Primary II | 2.5-3.4 |
| Primary III | 3.5-4.4 |
| Intermediate I | 4.5-5.4 |
| Intermediate II | 5.5-6.9 |

3. Subtest components

| | Primary I | Primary II | Primary III | Intermediate I | Intermediate II |
|---|---|---|---|---|---|
| Total Reading | | | | | |
| Reading Comprehension | X | X | X | X | X |
| Word Study Skills | X | X | X | X | X |
| Total Mathematics | | | | | |
| Concepts | X | X | X | X | X |
| Computation and Applications | X | | | | |
| Computation | | X | X | X | X |
| Applications | | X | X | X | X |
| Total Auditory | | | | | |
| Vocabulary | X | X | X | X | X |
| Listening Comprehension | X | X | X | X | X |

4. Norming: Empirical norms are available with a midpoint of 8 October for grades 2-9, and 8 May for grades 1-9, and 8 February for grades 1 and 2.

5. Out-of-level testing: Provides an expanded standard score scale. Testing more than one level out-of-level is not recommended.

6. Procedures for minimizing bias: Items were edited by a group of consultants with various minority backgrounds.

7. Other comments: Scaled score is continuous with Stanford Early School Achievement (SESAT) and Stanford Test of Academic Skills (TASK).

## TEST OF BASIC EXPERIENCE II
## (TOBE) 1978

1. Languages: English and Spanish

   The Spanish version is a direct translation from the English with the exception of items that would radically change in translation. In such cases equivalent items were constructed. Spanish version of the test occasionally provides a choice of words so that the most common version of words can be used with Mexican, Cuban, or Puerto Rican students.

2. Publisher's recommended in-level use:

| Level | Grade |
|---|---|
| K | Preschool, kindergarten, fall of first grade |
| L | Spring of kindergarten, first grade |

3. Subtests:

| | Level K | Level L |
|---|---|---|
| Mathematics | X | X |
| Language | X | X |
| Science | X | X |
| Social Studies | X | X |

4. Norming: Empirical norms exist only for the English version of the test; midpoints are October 19 and April 19.

5. Out-of-level testing: Provides expanded standard score scales.

6. Procedures for minimizing bias: Test items were reviewed by a panel of women and minority consultants. The Spanish version of the test was reviewed by native speakers of Puerto Rican, Cuban, and Mexican Spanish.

## Selecting a Language Proficiency Test

In order to select a language proficiency test, program personnel may consult catalogues of tests which are available. Some catalogues offer straight descriptions of instruments[1] while others offer an evaluative assessment of the tests.[2] It may be difficult to make a decision when confronted with so many choices of tests. In an effort to assist districts in this task, several states have convened panels of professionals with expertise in language proficiency testing for the purpose of examining and rating tests and making temporary recommendations.[3] The reports of such meetings are helpful to districts since they often explain the criteria upon which tests were selected and indicate the ratings given to each test.

---

[1]Bye, T. T. Tests that measure language ability: A descriptive compilation. Berkeley, California: BABEL/LAU Center, 1977.

Dissemination and Assessment Center for Bilingual Education. Evaluation instruments for bilingual education: An annotated bibliography. Austin, Texas: DACBE, 1976.

Northwest Regional Education Laboratory, Center for Bilingual Education. Assessment instruments in bilingual education: A descriptive catalogue of 342 oral and written tests. Los Angeles, California: National Dissemination and Assessment Center, 1978.

[2]Northwest Regional Educational Laboratory. Oral language tests for bilingual students: An evaluation of language dominance and proficiency instruments. Portland, Oregon: NWRL, 1976.

Pletcher, B. P., Locks, N. A., Reynolds, D. F., and Sisson, B. G. A guide to assessment instruments for limited English speaking students. New York, New York: Santillana Publishing Company, Inc. 1978.

[3]Law, A. Proceedings of the Bilingual Instrument Review Committee (AB 3470). Sacramento, California: Office of Program Evaluation and Research, California State Department of Education, September 28, 1978.

Texas Education Agency. Report from the committee for the evaluation of language assessment instruments, 1977.

The most important critical points to be taken into account in selecting a language proficiency test depend on the use to which the test will be put. Most districts use test results as the criterion (or one of the criteria) for classifying students as either limited in English or proficient in English. The validity of the test for this purpose, then, is of primary concern.

The test should provide a cutoff score or range and information about validity studies to support the cutoff. Unfortunately, at the time of printing, very few tests have adequate validity data and cutoff levels vary from test to test. This means that the same child might be classified as "limited-English-speaking" if Test A is administered and as "fluent-English-speaking" if Test B is administered. Studies are now being conducted to compare and equate language proficiency tests[4] and some helpful results should soon be available for making more informed decisions about using tests for program placement. Meanwhile caution should be exercised in relying on any single test for classifying students.

Another consideration related to validity concerns the scoring system. A test that has versions in two or more languages should provide a proficiency score in each language. It is illogical and inappropriate, however, to provide a proficiency rating in one language based only on proficiency in the other language. While a dominance classification can be derived from proficiency scores, a proficiency score cannot be determined on the basis of proficiency in the other language or on the basis of dominance.

It is difficult to get reliable, valid language proficiency scores for kindergarteners and first graders, particulary on the more global measures. One way to improve the situation is to be sure test administration procedures are strictly standardized and that children are not

---

[4]See for example, Gilmore, G., & Dickerson, A. The relationship between instruments used for identifying children of limited English speaking ability in Texas. Houston: Region IV Education Service Center,1979.

distracted. Children in this age range have a short attention span and may not be willing to sit still for 20 minutes. This problem might be overcome by administering a test in two parts. Since it is only possible to test certain aspects of language with any one test, and since valid reliable results are not assured, teacher judgment should play a part in arriving at decisions concerning classification and program placement.

If the test is to be used as an achievement measure, as well as a classification measure, as is the case in many programs, then several other issues should be considered.[5] First, the test needs to have enough items so that growth can be detected. Second, when children are tested every fall and spring with the same test, they may memorize parts of the test, particularly stories. Fall to fall testing would help, but then, whatever unknown amount of growth occurs during the summer cannot be attributed to the program. A third consideration concerns units for measuring growth. Some tests provide a score of one to five levels. Setting goals and reporting growth in this way masks growth that occurs within levels. A test should provide raw scores as well as levels, and growth should be reported in terms of raw scores. It may <u>also</u> be interesting to report changes in levels.

Here is a list of additional points to take into account in selecting and in interpreting the results from language proficiency tests:

1. Instructions should be simple and totally understandable to the student. They should be provided in the language the child knows best.

2. Administration procedures should be clearly spelled out so that they can be standardized across administrations.

---

[5]Using the same test for selection and pre-post outcome evaluation will introduce bias due to a regression toward the mean resulting in exaggerated gains (see Horst, Tallmadge, and Wood, 1975).

3. Elicitation tasks should not require unnatural language, the responses expected should be those of an average native speaker of the same age speaking in normal conversational style.

4. Items should not require tasks that are above the developmental level of the student.

5. Items should not require metalinguistic awareness or linguistic manipulation, since these may not be indicators of proficiency.

6. Items should measure aspects of language and not other things such as memory, literacy, and willingness to talk.

7. The content of the test should be within the student's cultural experience.

8. Proficiency should not be determined strictly on the basis of quantity of speech.

9. A test that is too long or too short may be unreliable.

## Annotated List of Language Proficiency Tests

This annotated list of language proficiency tests is short and provides project directors and evaluators with much of the information necessary to make a well informed choice. The criterion used for including tests in the list is the following: each test is recommended (at the time of printing) by at least one of the three states having the largest number of bilingual education programs. The tests are primarily in Spanish and English and range from kindergarten level to high school. A brief description is offered of each test as well as comments on the linguistic and technical properties of the tests. The comments are points that evaluators and project directors should be well aware of in selecting a test or in interpreting test results. The comments were drawn from several sources including the experience of districts in the bilingual PIP field test study, and published articles and critiques. Each publisher was given an opportunity to respond to the review and to include "Publisher's Comments." This information has been incorporated into the reviews.

## Descriptions of Commonly Used Language Proficiency Tests

### Basic Inventory of Natural Language (BINL)

| | |
|---|---|
| **Languages:** | English and Spanish (can be used for other languages) |
| **What It Tests:** | Speaking |
| **Levels and Grades:** | K-12 |
| **Administration:** | Individually administered. Requires 10-15 minutes.<br><br>Pictures are used to elicit natural speech and ten sentences are tape recorded for later analysis. |
| **Scoring:** | Hand or machine scored. |
| **Interpretation:** | Yields raw scores that can be converted to one of four levels: NES, LES, FES, PES ("proficient").<br><br>Age is taken into account in determining levels. |
| **Comments:** | Pictures are large, attractive, with multicultural content. It is difficult to standardize administration procedures since there is no set of "items" but rather an elicitation technique. Complex to score by hand. Scored on the basis of linguistic complexity and length of sentences. These criteria may not always be valid indicators of proficiency.<br><br>No information is provided on the validity of the proficiency categories. Information on validity is limited to correlations of sentence length with complexity, and correlations of complexity scores with an oral reading test. Reliability data is limited to correlations between the first half and the second half of the test. These correlations were high. Some districts have found that the test classifies fluent speakers as "limited" (see Gilmore and Dickerson, 1979). |
| **Publisher's Comment:** | Standardization is facilitated by adequate training and close adherence to BINL procedures. Machine scoring procedures: reports of five different types, from classroom listings to district summaries, including pre-post averages, minimum, maximum and average scores by grade levels. A recent study establishes averages for grades K-12 based on a sample of 125,000 students. Standard error allows for valid adjustment of scores. The format of the test permits retest on invalid tests which have been reported to be less than 4% of tests submitted for machine scoring. Percentile rank of scores is now included in reports. |

## Descriptions of Commonly Used Language Proficiency Tests

### Bilingual Syntax Measure (BSM)

| | |
|---|---|
| Languages: | English and Spanish |
| What It Tests: | Speaking |
| Levels and Grades: | Level I, K-2 (ages 4 to 9)<br>Level II (not available for review) |
| Administration: | Individually administered.<br>Requires 10-15 minutes.<br>Students respond orally to questions based on pictures. |
| Scoring: | Hand scored |
| Interpretation: | Provides language dominance (when both English and Spanish tests are administered), level of second language acquisition, and degree of maintenance or loss of the first language. Assigns students to one of five proficiency levels in each language. Additionally, provides instructional suggestions for reading and ESL which correspond to each of the five English proficiency levels. |
| Comments: | Attractive, colorful pictures are used to elicit speech through structured conversation. Responses are scored strictly on the correctness of specific grammatical structures. The choice of grammatical structures is based on research studies on the sequence of acquisition of morphemes. Allows for regional language variation. A number of discussions of this test have been published including Hernández-Ch., 1978[1] and Rosansky, 1979.[2]<br>Both test-retest reliability and inter-scorer reliability are reported in the Technical Handbook. Although the reported reliability is low, the authors attempt to explain why this is so (TH, p. 45). |

---

[1] Hernández-Ch., Eduardo. Critique of a critique: Issues in language assessment. Journal of the National Association for Bilingual Education, March 1978, Vol. II, No. 2.

[2] Rosansky, E. J. A review of the Bilingual Syntax Measure. In B. Spolsky, Advances in language testing, Arlington, VA: Center for Applied Linguistics, 1979.

## Descriptions of Commonly Used Language Proficiency Tests

### Comprehensive English Language Test for Speakers of English as a Second Language (CELT)

| | |
|---|---|
| **Language:** | English |
| **What It Tests:** | Listening comprehension, grammar, and vocabulary. Contains three subtests: (1) Listening, (2) Structure, and (3) Vocabulary |
| **Levels and Grades:** | High school, college, and adult.<br><br>Designed for intermediate to advanced ESL students. |
| **Administration:** | Group administered.<br><br>Listening requires 40 minutes; Structure requires 45 minutes; Vocabulary requires 35 minutes. A recording can be used to administer the listening test.<br><br>All test items are multiple choice. Students respond to oral and written stimuli by marking an answer sheet. |
| **Scoring:** | Scored with a key. |
| **Interpretation:** | Yields percent correct for each test.<br><br>Percentile scores are available (but see Comments).<br><br>Does not provide proficiency classifications. No cutoff score is provided for classification of students as limited in English proficiency, since test was not designed for this purpose. |
| **Comments:** | Oral production is not tested.<br><br>All test items on each subtest are multiple choice items that require reading; therefore, the measures of listening comprehension, structure, and vocabulary are each confounded with literacy skills. The authors recommend the Vocabulary subtest for use with students who have had advanced training in reading.<br><br>The three subtests had moderate to high internal consistencies with four groups of foreign students and, therefore, very reasonable standard |

Comprehensive English Language Test for Speakers of English as a Second Language (CELT) (continued)

| | |
|---|---|
| Comments: (continued) | errors of measurement. No information is given on predictive validity. Tentative evidence of concurrent validity is offered based on correlations with other standard ESL tests. Tentative norms for five different groups, based on small samples, are provided. The norms are not appropriate for use in most bilingual programs, however, since the students in the norming sample are not similar to most students in bilingual programs. |

## Descriptions of Commonly Used Language Proficiency Tests

### Ilyin Oral Interview Test

| | |
|---|---|
| Languages: | English |
| What It Tests: | Speaking |
| Levels and Grades: | Secondary and adult. |
| Forms: | There are two forms (BILL and TOM) and each has a long version (50 items) and a short version (30 items). |
| Administration: | Individually administered. Requires up to 30 minutes.<br><br>The students respond to pictorial stimuli and questions by responding orally. Items are ordered in difficulty and interview is terminated when a frustration level is reached. |
| Scoring: | Hand scored. |
| Interpretation: | Yields raw scores. No cutoff score is given to identify students as "limited" in English proficiency; however, suggestions are given for placement levels in adult ESL programs, and a range is suggested as the degree of proficiency required for jobs in which oral communication with the public is limited. |
| Comments: | The requirement to answer in a complete sentence is an unnatural one and may depress scores of students who fail to do this. The long version can become monotonous since many pictures are repeated.<br><br>Internal consistency reliabilities are high. No information is given for test-retest reliability or interrater reliability. Validity information is limited to correlations with other tests, and based on very small samples. |

## Descriptions of Commonly Used Language Proficiency Tests

### Language Assessment Battery (LAB)

| | |
|---|---|
| Languages: | English and Spanish |
| What It Tests: | Listening, speaking, reading, and writing.<br><br>Level I has three subtests: (1) Listening and Speaking, (2) Reading, and (3) Writing. Levels II and III have four subtests: (1) Listening, (2) Reading, (3) Writing, and (4) Speaking. |
| Levels and Grades: | Level I, grades K-2; Level II, grades 3-6; Level III, grades 7-12. |
| Administration: | Level I: Individually administered, requires 5-10 minutes.<br><br>Levels II and III: Part is individually administered; requires 41 minutes.<br><br>Students respond to verbal, written, and pictorial stimuli by pointing, by giving oral responses, by writing, and by marking answer sheets (on Levels II and III only). |
| Scoring: | Hand scored; parts scored with a key. |
| Interpretation: | Yields raw scores and stanines and percentiles by grade. Students scoring below the 20th percentile may be classified as limited in English proficiency. |
| Comments: | The speaking section of Level I, Test 1, contains only 6 items, all of which may be answered with one word. The writing tests measure reading skills in addition to writing skills.<br><br>The test went through all the stages of preparation by expert and experienced item writers, pilot studies, item- and test-analyses, and norming on substantial samples (20 schools, and about 500 students at each level from K through 12). The technical manual is a model.<br><br>One study[1] has shown that the Level I English test does not discriminate well in the range near the cutoff point for classifying students as limited in English. This reduces its value for use as a pre-post measure. |

[1] Hubert, J. An investigation of the Language Assessment Battery (English, Level 1) for Title VII students in Hartford. Unpublished manuscript, 1978.

## Descriptions of Commonly Used Language Proficiency Tests

### Language Assessment Scales (LAS)

| | |
|---|---|
| **Languages:** | English and Spanish |
| **What It Tests:** | Listening comprehension and speaking. Five subtests form the total score for both levels: (1) discrimination of minimal phonemic pairs, (2) vocabulary production, (3) phoneme production, (4) syntax comprehension, and (5) story production. |
| **Levels and Grades:** | Level I, grades K-5.<br>Level II, grades 6-12. |
| **Administration:** | Individually administered.<br><br>Requires 20 minutes.<br><br>Stimuli consist of tape recorded speech and pictures. Students respond orally, and by pointing. |
| **Scoring:** | Hand scored.<br><br>Interrater reliability should be obtained on storytelling task.<br><br>Age is taken into account in scoring. |
| **Interpretation:** | Yields a score of 1 to 100 which can be converted to a level, 1 to 5.<br><br>Students who score at level 3 or below are classified as "Limited English (or Spanish) speakers." |
| **Comments:** | This is a fairly comprehensive overall aural-oral proficiency test. There are problems with the phonemic discrimination section since this task requires a kind of metalinguistic awareness students may not have. The story retelling task measures not only production, but also comprehension.<br><br>Interrater reliability coefficients for the story retelling task are moderately high. Coefficients of internal item consistency for discrete-point items range from .36 to .96. |

Language Assessment Scales (LAS) (continued)

| | |
|---|---|
| Comments: (continued) | Validation consisted of one-way analyses of variance of relatively small samples (one- to two hundred) of students dichotomized into English-dominant and Spanish-dominant on the basis of teacher judgment.<br><br>Several studies of reliability were done on small samples (21 English and 35 Spanish) using various approaches. The sample sizes were too small to justify some of the analyses and the conclusions drawn from them. |

## Descriptions of Commonly Used Language Proficiency Tests

### Primary Acquisition of Language (PAL) Oral Language Dominance Measure (OLDM) Oral Language Proficiency Measure (OLPM)

| | |
|---|---|
| **Languages:** | English and Spanish |
| **What It Tests:** | Listening comprehension and speaking |
| **Levels and Grades:** | PAL OLDM, K-3<br>OLPM, 4-6 |
| **Administration:** | Individually administered.<br><br>Requires 15 minutes for each language.<br><br>Students respond orally to oral and pictorial stimuli. |
| **Scoring:** | Hand scored. |
| **Interpretation:** | Yields raw scores ("G scores") that are converted to proficiency levels, 1 to 5. Also yields dominance categories.<br><br>Students who score at level 4 or below are classified as "Limited English (or Spanish) speakers." |
| **Comments:** | Simple to use and score. Scored on the basis of grammaticality and appropriateness of responses as well as quantity of speech.<br><br>The test was developed "as a result of research by the El Paso Public Schools."<br><br>Item analyses were used in the construction of the tests although samples were somewhat small (about 200 drawn from three grades in high schools). Validity is quoted in terms of the tests ability to grade schools in correct order, and of correlations with a reading test. The latter were fair being around 0.3 to 0.5. |

## Descriptions of Commonly Used Language Proficiency Tests

### Shutt Primary Language Indicator Test (SPLIT)

| | |
|---|---|
| Languages: | English and Spanish |
| What It Tests: | Listening comprehension, speaking, reading, and grammar.<br><br>There are three subtests: (1) Listening Comprehension, (2) Verbal Fluency, and (3) Reading Comprehension and Grammar. |
| Levels and Grades: | Listening Comprehension, Verbal Fluency, K-6; Reading Comprehension and Grammar, 3-6. |
| Administration: | Listening Comprehension: Group administered; requires 35 minutes, tape recording available.<br><br>Verbal Fluency: Individually administered; requires 15 minutes.<br><br>Reading Comprehension and Grammar: Group administered; requires 30 minutes.<br><br>Instructions are provided in both languages and are available on tape. Stimuli are oral, pictorial, or written. Students respond orally, by marking pictures in answer book, or by marking an answer sheet. |
| Scoring: | Hand scored; parts scored with a key. |
| Interpretation: | Yields raw scores, percentile ranks, and age and grade equivalents.<br><br>Yields a dominance classification. |
| Comments: | Yields no cutoff point to classify students as limited in English proficiency (independent of Spanish/Portuguese score). A proficiency classification is given based on the dominance classification. This wrongly assumes that students are highly proficient in the dominant language. A student whose English score is very low can be classified as "English Adequate" if the student's Spanish score is also very low, but higher than the English score. Districts should establish their own cutoff points for classifying students in English.<br><br>Grade equivalent scores should not be used. |

## A System for Comparing Curriculum Content with the Content of CTBS Spanish and English, Form B and C

In order to measure program effects, the selection of a test that measures what is being taught is very important. Several systems have been developed to systematically compare curriculum content and test content.[1] Presumably, the evaluator will compare the program curriculum to several adequate and available tests and select the test that most closely matches the curriculum. The evaluator of a bilingual program has very few choices. At the time of this writing, the CTBS is the only widely used standardized achievement test battery that is available in both Spanish and English. Because this test is so widely used, a system of comparing its content to that of any curriculum would have wide application.

### Uses of the System

Test selection. As stated above, very few major comprehensive tests exist for the evaluation of bilingual programs. However, there are many locally developed tests that have been distributed and other tests that are fairly limited in scope. Also there are districts that chose to develop and use criterion referenced tests. Additional tests will undoubtedly be developed. There is the option with a test like the CTBS of using only the subtests that are appropriate or of testing out-of-level. Therefore, careful comparison of test content with curriculum content can be used to discard the CTBS if it is totally inappropriate or to select the best combination of subtests and/or the most appropriate levels.

Test interpretation. An evaluator may select the CTBS knowing that it does not match the curriculum as well as is desirable. A careful analysis of the test and the curriculum can still be a valuable tool for data analysis. The test items that match curriculum content can be analyzed

---

[1] Morris, L. L. and Fitz-Gibbon, C. T. "Determining How Well a Test Fits the Program" in How to Measure Achievement. Beverly Hills, California: Sage Publications, 1978.

separately from those which do not. If the gain for the matching items is greater than for the non-matching items, then a case can be made for program impact versus simple maturation.

Curriculum Planning. Another use of such a comparison is to make changes in the curriculum. This is not to suggest that "teach to the test" becomes the rule, because a test will always sample only a small amount of what is actually taught. However, curricula are always under revision and criteria by which success of instruction are to be evaluated have some claim to consideration.

## Limitations

This instrument has been developed only for the first two levels of the CTBS, levels B and C, commonly used in first and second grade. However, the steps outlined in this form could be used as a model for examining higher levels of the test.

## Directions for Use

The attached forms are divided into three parts per grade level: Spanish Reading, English Reading, and Math.

Each part consists of two sections: the Test/Curriculum Analysis, which is to be completed by each teacher; and the Summary, which is to be completed by the evaluator or other staff person. Where there are several teachers per grade level, the summary should represent an average. However, in cases where the instructional treatments varied so much that the test results will be reported separately, a summary should be made for each different treatment.

## Time for Task

Estimated working time is one hour per teacher to complete the analysis and several hours for the evaluator to explain the task to teachers, distribute and collect forms, and develop summaries.

CTBS English Reading Level B (Grade 1)

TEST/CURRICULUM ANALYSIS
(to be completed by project teachers)

Reading: Vocabulary from Tests 1, 2, and 3

1. As a result of the English language arts curriculum and other school and non-school experiences, what words on the word list are students likely to have seen, heard, read or used? Review the words on the word list and circle each word that the students have not been exposed to.

   CAUTION: A child knows many more words than are taught in school. Vocabulary is learned from many sources. Therefore, do not limit your consideration of students' vocabulary to what is covered in the curriculum.

## CTBS English Vocabulary -- Level B (all words from Tests 1-3)

a
after
and
animal
apple
are
around
at
balloon
beak
bed
big
Bill
Billy
birthday
bitten
black
Bob
book
boom
box
boy
breakfast
brown
brownie
bug
bunt
button
by
cake
came
can
cans
car
changed
children
chimney
Christmas
city
clamp
climb
clip
clock
clown
coat
dab
dollar
done
door
down
dress
drink
eggs
enamel
father
finger
fish
flower
fly
foot
for
Frank
frog
get
girl
girls
green
hand
happy
has
have
he
head
help
her
here
him
himself
hope
hot
I
in
into
is
it
jerk
Joan
jump
kitchen
kitten
know
lean
let
like
little
look
made
make
man
many
Mary
mender
mister
misters
money
mother
Mrs.
near
night
not
on
one
open
out
paint
party
people
pet
pig
plate
prince
puppy
rabbit
rain
read
ready
rope
safe
said
sat
school
schools
seal
see
sees
she
sheriff
show
sister
sisters
sleep
some
street
surprise
Susan
table
take
tell
the
these
they
this
to
took
toy
train
tree
truck
two
wagon
was
will
window
with
woman
won

Test 1: Word Recognition 1

Number of items: 19

Task: The student listens to a word read aloud and selects the correct printed word from four choices. Distractors consist of words that look similar to the right answer. Some are nonsense words or misspellings.

2. Have the students had practice reading English words up to three syllables long?

_____ daily or weekly
_____ only once or twice
_____ none

3. Have the students had practice reading all the letters and letter combinations that appear in Test 1? yes _____ no _____

If no, list letters or combinations that are not included in the curriculum: _____

Test 2: Reading Comprehension

Number of items: 24

Task: The student reads a sentence and selects an appropriate picture from three choices. Distractors consist of pictures with error in gender, error in number or error in content. About half of the items consist of two sentences; the other half consists of one sentence only. Sentences range from 3 to 10 words in length, with the average sentence having five words.

4. Have students had practice reading sentences in English?

_____ daily or weekly
_____ only once or twice
_____ none

Test 3: Word Recognition II

Number of items: 19

Task: The student chooses one of four printed words that best matches a picture. Twelve of the 19 words are identical to Test 1 Word Recognition I, but the tasks are different because in Test I students respond to an aural clue and in Test 3 to a visual clue.

No specific questions.

CTBS English Reading Level B (Grade 1)
(to be completed by project evaluator)

SUMMARY

Numbers in parentheses refer to question numbers on preceding pages.

1. What percent of the reading test vocabulary are students likely to have seen, heard, read or used? (1)

   _____% [The vocabulary list contains 167 different words.]

   Comments: _____

   _____

   _____

2. Have students been taught language arts skills tested? (2, 3, 4)

   Yes _____
   No _____

   Comments: _____

   _____

   _____

3. What major skills in the English language arts curriculum are not represented on the test?

   _____

   _____

4. What percentage of the curriculum does this represent?

   _____

   _____

CTBS Español Reading Level B (Grade 1)

TEST/CURRICULUM ANALYSIS
(to be completed by project teachers)

Reading: Test Vocabulary

1. As a result of the Spanish language arts curriculum and other school and non-school experiences, what words on the word list are students likely to have seen, heard, read or used? Review the words on the word list and circle each word that the students have not been exposed to.

   CAUTION: A child knows many more words than are taught in school. Vocabulary is learned from many sources. Therefore, do not limit your consideration of students' vocabulary to what is covered in the curriculum.



123

CTBS Español Vocabulary -- Level C (all words from Tests 1-3)

a
abajo
abiertas
abrir
agua
al
animal
animar
aquí
árbol
asomar
abril
ayudar
bajar
bajarse
beber
blanco
bocina
cabemos
cabeza
caja
calle
cama
camión
canción
Carlos
celos
cerdo
ciudad
clave
cocina
cochino
colina
color
come
comida
comprar
conejo
conoce
cuando
cuento
cuerda
cuidado
cumpleaños
de
dedo
dejó
dentro
dinero
disfrazó
dólar
dolor
dos
duro
el
ella
en
encima
enorme
es
esta
están
estas
este
falda
fiesta
flor
frota
fruta
fue
fuego
fuente
gato
gente
globo
gota
grande
guante
gusta
hay
hecho
hechos
hermana
hermanas
hermano
hermanos
hermoso
hermosos
hija
hijos
hizo
hojas
huerta
huevos
insecto
jota
Juan
juguete
la
las
latas
leer
les
libro
limpia
los
luego
Lupe
lleva
llora
lluvia
mamá
mano
mantel
mapa
María
me
minero
mira
mono
mosca
muchacho
muchachos
mujer
nación
negro
nieve
niña
niñas
niño
noche
nuevo
papá
para
paseo
pastel
pastor
payaso
peor
Pepe
pequeña
Pérez
perro
personas
pintar
plato
pobre
príncipe
pronto
puede
pueden
puedo
puerta
rama
rana
rata
ratón
reloj
rey
riño
ropa
sabían
salir
saltar
se
señor
señora
sentado
sentido
sillas
sólamente
solo
sombrero
son
sorpresa
su
sueño
Susana
tanta
tiene
tomo
trago
traje
tren
un
una
unas
usted
va
ve
vendido
venido
ventana
verdad
volvió
y
yo



124

Test 1: Reconocimiento de Palabras 1 (Word Recognition 1)

Number of items: 19

Task: The student listens to a word read aloud and selects the correct printed word from four choices. Distractors consist of words that look similar to the right answer. They might begin or end with the same sound, for instance.

2. Have the students had practice in Spanish reading words up to three syllables long?

_____ daily or weekly
_____ only once or twice
_____ none

3. Have the students had practice reading all the letters or combinations that appear in Test 1? yes _____ no _____

If no, list letters or combinations that are not included in the curriculum: _____

Test 2: Comprehensión de Lectura (Reading Comprehension)

Number of items: 24

Task: The student must read a sentence and select an appropriate picture from three choices. Distractors consist of pictures with error in gender, error in number or error in content. About half of the items consist of two sentences. The other half consists of one sentence only. Sentences range from 3 to 12 words in length, with the average sentence having five words.

4. Have students had practice reading sentences in Spanish?

_____ daily or weekly
_____ only once or twice
_____ none

Test 3: Reconocimiento de Palabras II (Word Recognition II)

Number of items: 19

Task: The student chooses one of four printed words that best matches a picture. Twelve of the 19 words are identical to Test 1 Word Recognition I, but the tasks are different because in Test I students respond to an aural clue and in Test 3 to a visual clue.

No specific questions.

CTBS Español Reading Level B (Grade 1)
(to be completed by project evaluator)

SUMMARY

Numbers in parentheses refer to question numbers on preceding pages.

1. What percent of the reading test vocabulary are students likely to have seen, heard, read or used? (1)

   _____% [The vocabulary list contains 197 words.]

   Comments: _____

   _____

   _____

2. Students have been taught language arts skills tested. (2, 3, 4)

   Yes _____
   No _____

   Comments: _____

   _____

   _____

3. What major skills in the Spanish language arts curriculum are not represented by the test?

   _____

   _____

4. What percentage of the curriculum does this represent?

   _____

CTBS Spanish or English Math Level B (Grade 1)

TEST/CURRICULUM ANALYSIS
(to be completed by project teachers)

Math Battery

1. What percent of math curriculum is devoted to computations? _____

2. What percent of math curriculum is devoted to math concepts, applications, and story problems? _____

3. Do students have adequate vocabulary in the language in which they are tested so they understand all directions and word problems? _____ yes _____ no

Test 4: Conceptos y Aplicaciones de Matemáticas/Mathematics Concepts and Applications

Number of items: 24

Task: The student listens to a problem or a question read aloud and selects from four possible answers.

4. Following is a list of the skills included in this test, with the number of items devoted to each skill listed in parenthesis. Check in the space provided whether each skill is covered in the curriculum and decide how many total items this represents.

| | Yes | No |
|---|---|---|
| value of numbers (2) | _____ | _____ |
| addition and subtraction (4) | _____ | _____ |
| numeration (3) | _____ | _____ |
| equating a set to a number (1) | _____ | _____ |
| equating a set to a number word (1) | _____ | _____ |
| counting by twos (1) | _____ | _____ |
| sets (1) | _____ | _____ |
| subtraction story problem (2) | _____ | _____ |
| missing addend (2) | _____ | _____ |
| setting up story problems for addition (1) | _____ | _____ |
| telling time (2) | _____ | _____ |
| measurement (1) | _____ | _____ |
| value of money (3) | _____ | _____ |

_____ of 24 items represent skills covered in the curriculum.

(Caution: Do not simply add checks. For each item checked add the number in the parenthesis at the end of that line.)

Test 5: Computación de Matemáticas/Mathematics Computation

Number of items: 32

Task: The student computes written addition problems and chooses the correct answer from a group of three. A page of subtractions, also with three possible answer choices, follows. The time allotted to this subtest averages one minute per computation.

5. What percent of the computations in math curriculum are represented by:

| | |
|---|---|
| addition | _____ |
| subtraction | _____ |
| total | 100% |

6. What percent of the additions performed in the math curriculum are represented by:

| | |
|---|---|
| horizontal addition | _____ |
| vertical addition | _____ |
| total | 100% |

| | |
|---|---|
| one digit addition | _____ |
| two digit addition | _____ |
| total | 100% |

7. What percent of the subtractions performed in the math curriculum are represented by:

| | |
|---|---|
| horizontal subtraction | _____ |
| vertical subtraction | _____ |
| total | 100% |

| | |
|---|---|
| one digit subtraction | _____ |
| two digit subtraction | _____ |
| total | 100% |

| | |
|---|---|
| subtractions requiring borrowing | _____ |

CTBS Spanish or English Math Level B (Grade 1)

SUMMARY
(to be completed by project evaluator)

Numbers in parentheses refer to question numbers on preceding pages.

1. Compare curriculum to test. (1, 2)

| | Percent of Curriculum | Percent of Test |
|---|---|---|
| Computations | _____ | 57 |
| Math concepts, applications, story problems | _____ | 43 |

Match is appropriate? yes ____ no ____

Comments: _____

_____

_____

2. Students have an adequate vocabulary for the math test? (3)

Yes _____
No _____

Comments: _____

_____

_____

3. In the math concept test, _____ out of 24 or _____% of the test represents items that students have practiced in the curriculum. (4)

Comments: _____

_____

_____

4. Compare curriculum to test. (5, 6, 7)

| | Percent of Curriculum | Percent of Test |
|---|---|---|
| addition | _____ | 50 |
| subtraction | _____ | 50 |
| horizontal addition | _____ | 31 |
| vertical addition | _____ | 69 |
| one digit addition | _____ | 31 |
| two digit addition | _____ | 69 |
| horizontal subtraction | _____ | 31 |
| vertical subtraction | _____ | 69 |
| one digit subtraction | _____ | 6 |
| two digit subtraction | _____ | 94 |
| subtraction with borrowing | _____ | 6 |

Math computation skills are represented in the curriculum in similar proportion to their appearance on the test?

Yes _____
No _____

Comments: _____

_____

_____

5. What skills from the math curriculum are not represented by the test?

_____

6. What percentage of the curriculum does this represent?

_____

CTBS English Reading Level C (Grade 2)

TEST/CURRICULUM ANALYSIS
(to be completed by project teachers)

**Reading: Test Vocabulary**

1. As a result of the English language arts curriculum and other school and non-school experiences, what words on the word list are students likely to have seen, heard, read or used? Review the words on the word list and circle each word that the students have not been exposed to.

   CAUTION: A child knows many more words than are taught in school. Vocabulary is learned from many sources. Therefore, do not limit your consideration to what is covered in the curriculum.

## CTBS English Vocabulary -- Level C (all words from Tests 1-3)

a
above
admire
act
afraid
after
against
age
air
all
almost
am
and
Ann's
are
arrive
around
as
astronauts
at
automobile
awake
away
awful
baby
back
bake
ball
bark
barked
barn
batter
be
beach
beans
began
begin
Betty
bicycle
birthday
blank
block
boat
boil
bold
books
boy
boys
box
brave
bread
bright
broken
brought
bug
buy
came
can
candy
cannot
car
care
careful
cat
catch
caught
children
city
clap
clean
clock
coat
comb
come
coming
cookies
cost
could
count
cow
cowboys
crawl
crying
curly
dad
danger
daughter
day
dear
different
did
dig
dime
dirty
dish
dog
down
dry
dull
during
earth
egg
end
enjoy
everyone
except
face
family
far
farm
father
feather
feet
fell
few
find
fish
foolish
foot
for
forest
forget
form
found
four
Friday
friend
friends
frisky
front
fry
full
fun
fur
gallop
game
gay
get
girls
give
glass
gloves
go
going
good
got
grandmother
green
grew
growing
had
hair
hall
hand
happy
hard
has
hate
hats
have
having
he
head
heavy
help
her
here
hide
high
hill
hills
him
his
hole
hollow
home
hop
horse
hot
hotel
hour
house
hungry
hurt
I
ice cream
if
in
Indian
Indians
is
it
Jack
Jill
Jim
Joe
jumped
Kathy
kicked
kind
kitten
know
lady
lag
larger
left
letter
let
let's
library
light
like
lion
listen
little
live
load
loud
low
make
man
map
me
men
milk
miss
moccasins
mode
moon
morning

Vocabulary List (Continued)

most
mother
Mr.
music
must
my
name
near
need
new
nice
nine
noises
not
November
now
o'clock
old
of
on
one
only
orange
orbit
our
outside
oven
over
package
paid
pans
paper
parks
party
Patty
people
picture
pigs
pillow
pink
pitch
plate
play
plays
please
pointing
pony
poor
pretty
put
ran
rather
regular
rent
rich
ride
ring
river
rocket
rocks
room
running
sad
saddle
said
sail
sail boat
Sam
same
sand
saw
say
school
sea
seat
see
seemed
sell
sent
she
shells
shiny
shoes
should
shout
shovel
show
silver
sister
skipper
sleepy
slept
slide
small
smaller
smart
smile
Smith
so
soft
some
somebody
son
song
soon
soup
spacemen
spell
spill
spoon
stars
stayed
stones
store
street
strong
stuck
sun
supper
surprise
take
teach
thank
that
the
they
them
then
think
this
three
time
to
today
Tommy
tomorrow
too
toward
toy
train
tree
trick
tried
trip
truck
two
ugly
uncle
under
unpainted
until
up
us
very
visit
wagon
wait
waiting
walk
want
was
washed
water
we
week
went
were
what
wheel
when
where
while
white
who
why
wide
wig
wildcat
will
wise
wish
with
wore
work
worker
worn
yeard
yet
you
young
your

Test 1 -- Reading Vocabulary

Number of items: 33

Task: The student listens to the definition of a word read aloud. For each item, the student selects from four printed words the one that best fits the definition. Distractors include antonyms, contextually related words, and unrelated words.

2. Have the students had practice in supplying a word in English to fit a definition?

_____ yes, using a format identical to test items
_____ yes, but using another format
_____ no

Test 2: Reading Comprehension: Sentences

Number of items: 23

Task: The student reads a sentence and selects the word that best completes the sentence. A block of four answer choices is offered and is located at the point in the sentence where the word is missing: initial position, medial position, or final position. The sentence completion item most often occurs in the middle of the sentence. The average sentence length is seven words.

3. Have the students had practice reading complete sentences in English of at least seven words in length?

_____ daily or weekly
_____ only once or twice
_____ none

4. Have the students had practice supplying a missing word in a sentence?

_____ yes, using a format identical to test items
_____ yes, but using another format
_____ no

## Test 3: Reading Comprehension: Passages

Number of items: 18

Task: The student reads six brief passages. Each passage is followed by two to four multiple choice questions to be answered by the student. These questions involve literal and near literal recall, use of context clues, stating main ideas, drawing conclusions, and recalling sequence. Paragraphs range from 5 to 14 sentences in length. The average sentence is 8 words long.

5. Have the students had practice reading paragraphs in English that are at least 5 sentences in length?

_____ daily or weekly
_____ only once or twice
_____ none

6. Have students had practice in answering questions based on reading paragraphs in English?

_____ daily or weekly
_____ only once or twice
_____ none

7. If students have had such practice, what percentage of classroom questions based on reading paragraphs utilize the following skills?

| | less than 20% | between 20% and 50% | more than 50% |
|---|---|---|---|
| literal and near literal recall | _____ | _____ | _____ |
| use of context clues | _____ | _____ | _____ |
| stating main ideas | _____ | _____ | _____ |
| drawing conclusions | _____ | _____ | _____ |
| recalling sequence | _____ | _____ | _____ |
| other | _____ | _____ | _____ |

CTBS English Reading Level C (Grade 2)

SUMMARY
(to be completed by project evaluator)

Numbers in parentheses refer to question numbers on preceding pages.

1. What percent of the reading test vocabulary are students likely to have seen, heard, read or used? (1) _____% (The vocabulary list contains 421 words.)

   Comments: _____

   _____

2. The language arts skills that are tested are also part of the curriculum. (2, 3, 4, 5, 6) Yes _____ No _____

   Comments: _____

   _____

3. Compare the kinds of questions asked in the reading test to the kinds of questions asked in the reading curriculum. (7)

| | Percent of Curriculum | Percent of Test |
|---|---|---|
| literal and near literal recall | _____ | 50 |
| use of context clues | _____ | 11 |
| stating main ideas | _____ | 5.5 |
| drawing conclusions | _____ | 28 |
| recalling sequence | _____ | 5.5 |
| other | _____ | 0 |

   The kinds of questions asked in the reading portion of the test are also practiced in the reading curriculum in a fairly similar proportion. Yes _____ No _____

   Comments: _____

   _____

4. What major skills in the English language arts curriculum are not represented in the test?

   _____

   _____

5. What percentage of the curriculum does this represent?

   _____

CTBS Español Reading Level C (Grade 2)

TEST/CURRICULUM ANALYSIS
(to be completed by project teachers)

**Reading: Test Vocabulary**

1. As a result of the Spanish language arts curriculum and other school and non-school experiences, what words on the word list are students likely to have seen, heard, read or used? Review the words on the word list and circle each word that the students have not been exposed to.

   CAUTION: A child knows many more words than are taught in school. Vocabulary is learned from many sources. Therefore do not limit your consideration of students' vocabulary to what is covered in the curriculum.

## CTBS Español Vocabulary -- Level C (all words from Tests 1-3)

a
abajo
abuelita
acto
admirar
afortunado
afuera
agarrar
agua
ahora
aire
al
alegre
algo
alguien
algunas
algunos
almohada
alrededor
alto
alquiler
amigo
amigos
Ana
ancho
anillo
apellido
aplaude
aquí
árbol
arena
arriba
asiento
astronautas
atrás
atrevido
autobús
automóvil
avión
avisarle
avisennos
ayuda
ayudar
ayudarlo
barco
bailar
bateador
Beatriz
bebé
biblioteca
bicicleta
blanca
Blanco
blancos
bonitas
bonitos
bosque
brazo
brillante
brillantes
brinca
bueno
caballito
caballo
cabeza
caer
caja
calor
caliente
callado
calle
caminar
camión
camisa
canción
cargar
cara
carro
carta
casa
casi
Cata
cavar
cayó
cena
cerca
cerro
ciudad
cochinitos
cogió
cohete
comenzó
comer
comió
como
comprar
con
conchitas
contar
contra
corriendo
corrió
costo
creer
cuál
cuando
cuatro
cuchara
cuento
cuidado
cumpleaños
Chávez
chica
chico
dar
de
debajo
debe
debería
debo
dejar
del
deletrear
desear
despierto
despintado
después
día
dibujo
dice
diez
diferente
dijo
divirtió
divirtiendo
dónde
dormido
dos
dulces
dura
durmió
duro
e
edad
el
ella
ellos
empezar
en
encontramos
encontraron
enojó
ensalada
enseñaré
enseñar
entre
era
eran
es
ese
esa
esconder
escondieron
escuchen
escuela
esperamos
esperen
esta
estaba
estaban
establo
estamos
este
esto
estoy
estrellas
estuvieras
excepto
familia
favor
felices
feliz
feo
fiesta
frijoles
fuerte
fuertes
fue
galopar
galletas
García
gatear
gatito
gato
gente
gracias
grande
gritar
guantes
gusta
gustan
hablo
hacer
hacerse
hacía
hambre
hambriento
hasta
helado
hermana
hermano
hermoso
hija
hijo
hizo
hombre
hombres
hora
horno
horrible
hotel
hoy
hoyo
hueco
huevo
iba
iban
iglesia
indio
indios
ir
irme
irse
Jaime
jardín
Jesús
José
joven
juegan
jugamos
jugar
juguete

Vocabulary List (Continued)

juguetes
la
ladrar
ladró
las
lastimó
lavado
le
leche
lejos
león
libras
ligero
limpio
listo
Lobo
lodosos
los
López
luego
luna
llama
llamar
llegar
lleno
llevaba
llorando
madre
mamá
mañana
mano
mapa
mar
María
marinero
más
me
medio
mejor
metió
mi
miedo
miedoso
mientras
mismo
mocasines
mojado
monte
montés
montura
mucho
música
muy
necesita
negros
niñas
niño
niños
no
nombre
nosotros
noviembre
nueve
nuevos
obra
obscura
obscuro
odiar
oigan
olvidar
oración
órbita
otra
otras
padre
pagoda
pala
palabra
pan
papá
papel
paquete
para
parecía
pasaba
paseo
pasó
pastel
pata
pateó
Patricia
peine
peligro
pelota
peluca
pensar
pequeña
pequeño
Pérez
periódico
pero
perro
pesado
pesados
pescados
piedras
piel
pierna
pies
pintando
plata
plato
playa
pluma
pobre
pocos
podían
poner
por
primero
programa
pueden
punto
puso
que
quedamos
quedan
querido
queridos
quien
quienes
quisiera
rancho
rato
regalo
regresamos
regular
reloj
resbalar
rico
río
rizada
rocas
rojos
rosadas
rota
roto
rueda
ruidos
sabio
salón
saltó
se
seco
semana
señor
señora
señorita
será
si
simpático
sobre
sol
sólo
sombrero
sombreros
son
sonríen
sonrisa
sonrió
sopa
sorpresa
su
suave
sucio
sucios
sus
también
te
teatro
ten
tenía
tiempo
tiene
tienda
tierra
tío
tipo
tirar
todavía
todo
todos
toma
tomar
Tomás
tonto
trabajador
trabajar
trajo
trató
tren
tres
triste
tristes
truco
tu
un
una
unas
uno
unos
va
vaca
valiente
valor
vamos
van
vaqueros
vaso
venado
vender
vengan
venir
ver
viajaban
viajamos
viaje
viejo
viejos
viene
viernes
vinieron
vino
vió
visitar
vive
voy
y
yo
zapatos

Test 1 -- Vocabulario de Lectura (Reading Vocabulary)

Number of items: 33

Task: The student listens to the definition of a word read aloud. For each item, the student selects from four printed words the one that best fits the definition. Distractors include antonyms, contextually related words, and unrelated words.

2. Have the students had practice in supplying a word in Spanish to fit a definition?

_____ yes, using a format identical to test items
_____ yes, but using another format
_____ no

Test 2: Comprensión de Lectura: Oraciones (Reading Comprehension: Sentences)

Number of items: 23

Task: The student reads a sentence and selects the word that best completes the sentence. A block of four answer choices is offered and is located at the point in the sentence where the word is missing: initial position, medial position, or final position. The sentence completion item most often occurs in the middle of the sentence. The average sentence length is seven words.

3. Have the students had practice reading complete sentences in Spanish of at least seven words in length?

_____ daily or weekly
_____ only once or twice
_____ none

4. Have the students had practice supplying a missing word in a sentence?

_____ yes, using identical format as test items
_____ yes, but using another format
_____ no

**Test 3: Comprensión de Lectura: Pasajes (Reading Comprehension: Passages)**

**Number of items:** 18

**Task:** The student reads six brief passages. Each passage is followed by two to four multiple choice questions to be answered by the student. These questions involve literal and near literal recall, use of context clues, stating main ideas, drawing conclusions, and recalling sequence. Paragraphs range from 5 to 17 sentences in length. The average sentence is 8 words long.

5. Have the students had practice reading paragraphs in Spanish that are at least 5 sentences in length?

   _____ daily or weekly
   _____ only once or twice
   _____ none

6. Have students had practice in answering questions based on reading paragraphs in Spanish?

   _____ daily or weekly
   _____ only once or twice
   _____ none

7. If students have had such practice, what percentage of classroom questions based on reading paragraphs utilize the following skills?

| | less than 20% | between 20% and 50% | more than 50% |
|---|---|---|---|
| literal and near literal recall | _____ | _____ | _____ |
| use of context clues | _____ | _____ | _____ |
| stating main ideas | _____ | _____ | _____ |
| drawing conclusions | _____ | _____ | _____ |
| recalling sequence | _____ | _____ | _____ |
| other | _____ | _____ | _____ |

CTBS Español Reading Level C (Grade 2)

SUMMARY
(to be completed by project evaluator)

Numbers in parentheses refer to question numbers on preceding pages.

1. What percent of the reading test vocabulary are students likely to have seen, heard, read or used? (1) _____% (The vocabulary list contains 482 words.)

   Comments: _____

   _____

2. The language arts skills that are tested are also part of the curriculum. (2, 3, 4, 5, 6) Yes _____ No _____

   Comments: _____

   _____

3. Compare the kinds of questions asked in the reading test to the kinds of questions asked in the reading curriculum. (7)

| | Percent of Curriculum | Percent of Test |
|---|---|---|
| literal and near literal recall | _____ | 50 |
| use of context clues | _____ | 11 |
| stating main ideas | _____ | 5.5 |
| drawing conclusions | _____ | 28 |
| recalling sequence | _____ | 5.5 |
| other | _____ | 0 |

   The kinds of questions asked in the reading portion of the test are also practiced in the reading curriculum in a fairly similar proportion. Yes _____ No _____

   Comments: _____

   _____

4. What major skills in the Spanish language arts curriculum are not represented in the test?

   _____

   _____

5. What percentage of the curriculum does this represent?

   _____

CTBS Spanish or English Math Level C (Grade 2)

TEST/CURRICULUM ANALYSIS
(to be completed by project teachers)

Math Battery

1. What percent of math curriculum is devoted to computations? _____

2. What percent of math curriculum is devoted to math concepts, applications, and story problems? _____

3. Do students have adequate vocabulary in the language in which they are tested so they understand directions and word problems?
yes _____ no _____

Test 4: Computación de Matemáticas/Mathematics Computation

Number of items: 28

Task: The student performs a computation and chooses the correct answer from the four that are provided. The computations consist of addition, subtraction, and multiplication.

4. What percent of the computations in math curriculum are represented by:

| | |
|---|---|
| addition | _____ |
| subtraction | _____ |
| multiplication | _____ |
| total | 100% |

5. What percent of the additions performed in the math curriculum are represented by:

| | |
|---|---|
| horizontal addition | _____ |
| vertical addition | _____ |
| total | 100% |
| one-digit addition | _____ |
| two-digit addition | _____ |
| three-digit addition | _____ |
| total | 100% |
| additions requiring carrying | _____ |
| additions with decimals | _____ |

6. What percent of the subtractions performed in the math curriculum are represented by:

| | |
|---|---|
| horizontal subtraction | _____ |
| vertical subtraction | _____ |
| total | 100% |

one-digit subtraction _____
two-digit subtraction _____
three-digit subtraction _____
total 100%

7. What percent of the multiplications performed in the math curriculum are represented by:

horizontal multiplication _____
vertical multiplication _____
total 100%

one digit multiplication _____

Test 5: Conceptos y Aplicaciones de Matemáticas/Mathematics Concepts and Applications

Number of items: 25

Task: The student listens to a problem or a question read aloud and selects from four possible answers.

8. Following is a list of the skills included in this test, with the number of items devoted to each skill noted in parenthesis. Check in the space provided whether each skill is covered in the curriculum, and decide how many total items this represents.

_____ of 25 items represent skills covered in the curriculum.

| | Yes | No |
|---|---|---|
| addition story problem (4) | ____ | ____ |
| equating number word to a set of items (1) | ____ | ____ |
| counting by more than 1 (2) | ____ | ____ |
| liquid measures (1) | ____ | ____ |
| adding money (2) | ____ | ____ |
| subtracting money (1) | ____ | ____ |
| applied numeration (days of week) (1) | ____ | ____ |
| telling time (2) | ____ | ____ |
| simple fractions (3) | ____ | ____ |
| single digit horizontal addition, addends precede sum (1) | ____ | ____ |
| single digit horizontal addition, sum precedes addend (1) | ____ | ____ |
| setting up a story problem for addition (1) | ____ | ____ |
| setting up a story problem for subtraction (1) | ____ | ____ |
| application of addition to a ruler-like scale (1) | ____ | ____ |
| application of addition to time (1) | ____ | ____ |
| missing subtrahend (1) | ____ | ____ |

(Caution: do not simply add checks. For each item checked add the number in parenthesis at the end of that line.)

CTBS Spanish or English Math Level C (Grade 2)

SUMMARY
(to be completed by project evaluator)

Numbers in parentheses refer to question numbers on preceding pages.

1. Compare curriculum to test. (1, 2)

| | Percent of Curriculum | Percent of Test |
|---|---|---|
| Computations | _____ | 53 |
| Math concepts, applications, story problems | _____ | 47 |

Match is appropriate? yes ____ no ____

Comments: _____

_____

_____

2. Students have an adequate vocabulary for the math test? (3)

Yes _____
No _____

Comments: _____

_____

_____

3. Compare curriculum to test. (4, 5, 6, 7)

| | Percent of Curriculum | Percent of Test |
|---|---|---|
| addition | | 36 |
| subtraction | | 36 |
| multiplication | | 28 |
| total | 100 | 100 |
| horizontal addition | | 40 |
| vertical addition | | 60 |
| total | 100 | 100 |
| one digit addition | | 20 |
| two digit addition | | 60 |
| three digit addition | | 20 |
| total | 100 | 100 |
| addition requiring carrying | | 40 |
| addition with decimals | | 10 |
| horizontal subtraction | | 40 |
| vertical subtraction | | 60 |
| total | 100 | 100 |
| one digit subtraction | | 0 |
| two digit subtraction | | 80 |
| three digit subtraction | | 20 |
| total | 100 | 100 |
| horizontal multiplication | | 100 |
| vertical multiplication | | 0 |
| total | 100 | 100 |
| one digit multiplication | | 100 |

4. In the math concept test, ____ out of 25 or _____% of the test represents items that the students have practiced in the curriculum. (8)

Comments: _____

_____

5. What major skills from the math curriculum are not represented by the test?

_____

_____

6. What percentage of the curriculum does this represent?

_____

_____



147

# COLLECTING DATA

MAJOR CONTENT ITEMS

7-A. DATA COLLECTION PROCEDURES (CHECKLIST)

7-B. DATA RECORDING FORM

SECTION 7

# 7. COLLECTING DATA

Data collection includes obtaining student background information, gathering teacher opinions, observing classroom operation, and a variety of other activities, but the focus of this section is the administration and scoring of tests and the recording of the scores. Of all the topics addressed in this manual, data collection is the only one with no major, unresolved theoretical issues. To obtain clean data, all that is required is to follow simple, widely known procedures. Yet data collection problems are a major reason for the lack of credibility in educational evaluations.

Key Problems

Testing procedures. Adequate testing procedures simply require following the publisher's instructions exactly, and making sure that pre- and posttesting conditions and procedures are identical. While this is not difficult, it does require some effort on everyone's part. Most problems are probably due to a lack of understanding of the importance of careful data collection. See Item 7-A.

Test scoring and data recording. Both scoring and recording are subject to clerical errors, but these errors can easily be held to an acceptable level through adequate care and accuracy checks. More difficult to deal with are scoring procedures that require the scorer to make judgments. (See Item 7-A.) The major problems in recording data are to provide all the essential information in a manageable format. (See Item 7-B).

Related Issues

Training testers. For experienced testers using a familiar test it is sufficient to bring the group together briefly within a few days of the beginning of testing to review the tests and testing procedures. For new tests or inexperienced testers, each tester must practice administering the entire test under the supervision of the evaluator.

Testing on appropriate dates. Testing should be done within a few days of the same date each year. For norm-referenced evaluations the testing should be within a week or two of the time that normative data were collected by the test publisher (or local district). Tests must also be spread out over days so that the burden on the students is not so great as to lower scores. Pre- and posttesting must follow similar schedules.

Recording data for longitudinal evaluations. A data recording form that works well for a single fall-to-spring evaluation may not be suitable for following student progress over several years. Student attrition, regrouping of classes each year, and the total number of scores involved all present problems. Appropriate individual student record forms may be the best solution. See Item 7-B.

## Data Collection Procedures
Outline

1. Assembling the students

   - Similar testing conditions for all treatment and comparison groups should be utilized. The time, place, and date of test administration should be considered. Technical manuals for test administration often contain testing procedure recommendations (i.e., avoid afternoon testing, or testing on Monday and Friday).

   - Distractions should be minimized. Avoid testing in the hall, or in the cafeteria as lunch is being prepared.

   - Coordinate testing efforts with district testing or assessment policies and procedures.

   - Consider teaching test-taking skills to students. This includes acquainting students with test formats, etc., NOT teaching to the actual test.

   - Plan for make-up testing

2. Administering the test

   - Identify testers. If teachers do not speak the appropriate language, identify alternative testers.

   - Conduct inservice training for all test administrations. If aides and parents will be used in testing, more intensive training will be required for them. The items on the list below should be addressed:
     - Familiarity with materials
     - Clarity of presentation
     - Adherence to guidelines and time limits
     - Control in the classroom
     - Attention to physical conditions (e.g., seat spacing)
     - Practice for individual testing
     - Correct choice of testing dates (e.g., norming dates)
     - The need for the inevitable "fill-in" of absentees

   - Clearly define roles and responsibilities of testors. Inservice training and determination of roles and responsibilities should be assertively coordinated by the project director.

3. Scoring the test

   - Train test scorers.
   - Scored tests should be spot checked by someone else.
   - Check interrater reliability.

4. Recording scores

   (See Item 7-B)

## Data Recording Forms

Recording the scores is the final step in the data collection process but, to ensure that the scores will be usable, the details of recording should be worked out well before pretest time. Where a commercial scoring service is used, the school evaluator may have little control over the recording process, but if the school elects to do its own scoring or wishes to transfer scores from computer printouts to a more convenient form, the evaluator must consider two important issues: the accuracy of the data, and the details of the data recording forms.

Copying scores accurately onto data forms is not a complicated problem for small-scale local studies, but it must not be overlooked. Even the most conscientious recorders make errors, and all data forms should be carefully proofread, preferably with one person reading aloud while a second person checks the scores.

The details of the data forms might appear to be of little importance, but in many school districts the way in which data have been recorded virtually precludes any reasonable analyses. It is not possible to prescribe a standard data format because school requirements vary so widely, but it is possible to state two general principles which must be observed. First, all scores must be completely identified, and second, scores must be arranged in a way that facilitates analysis. Sample data forms illustrating these principals are attached. Specific issues related to the use of such forms are discussed below.

Considerations for data recording forms

1. Most sets of scores require more than one page. The page number identifies each sheet and the "number of pages" helps make sure no pages are missing.

2. Every sheet of paper should have a name and date to indicate who filled in the numbers in case any questions arise in the future.

3. The group for which data are recorded should be clearly identified at the top of the page to simplify the retrieval of that group's data from a large data base.

4. The page should be arranged so that it can be photocopied without the students' names. This permits wide use of the data for research purposes without compromising student privacy.

5. It simplifies analysis greatly to have only one test (pre and post) recorded on each sheet, provided the rules for listing students (see points 6-11 below) are followed. The complete name of the pretest and posttest (taken exactly from the test booklets and including publication date) must be listed. This point is widely neglected.

6. Identifying students and organizing their names efficiently are the most difficult problems in recording student data. Where evaluations are only for one year and are based on fall and spring testing, the problems can be solved with a little effort and care. But where students must be followed over several years, there is no simple solution since students come and go from projects, and groups are reorganized every year. The simplest rule is to make sure that the posttest scores are all entered on the same sheet of paper as the corresponding pretest scores. This at least eliminates the problem of the evaluator trying to find each student's name on two lists.

7. A second rule for listing student names is to establish a standard ordering of the names, and stick to it for the life

of the evaluation and for all tests that are used. If a student moves or fails to take some of the tests, then the appropriate entries are blank, of course, but he or she should not be eliminated from the list. If new students enter the program, their names should be added to the end of the lists for all tests, even those for which no data will be entered. In addition to the obvious reduction in confusion, there are some practical advantages to this procedure. For example, a master form can be prepared with only the students' names and identification numbers filled in, and the forms can simply be duplicated when new tests are given. It also makes comparisons or correlations between any two sets of scores relatively easy because any two forms can be laid side by side and the corresponding names will line up correctly. If there is a compelling reason to change the order of student names in the middle of a project, then either all forms should be changed, or a double set of forms (old and new order) should be maintained.

8. A rule should be established for recording names. "Caldwell, D. E." should never become "Danny Caldwell" on a second list. The simplest procedure is to allow plenty of space and to spell out first names and middle initials (e.g., Caldwell, Daniel E.).

9. Each student should have an ID number that completely identifies him or her. The example in Figure 4 uses a one-digit experimental condition number, a two-digit group or class identification, a one-digit sex code, and a two-digit student number. In some evaluations, other codes (including letters) can be used, but careful consideration of the situation is necessary in order to permit any desired grouping simply by ID number.

10. A page should have some reasonable number of entries, probably 20 or 25. For some inexplicable reason, numbers like 27 and 33 are popular, and often the number of entries varies from page to page. Unnecessary complications like this help to make the statistician's life miserable.

11. Test dates are critical, especially in norm-referenced evaluations. If all students listed on a page have their pretests in one day and all are later posttested in a single day, then the test date column is not really necessary. However, this is usually impossible to predict at the time the form is made up, so the columns should be there in order to permit identification of make-up tests and late entries into the program.

12. Pre- and posttest scores should, in general, be in adjacent columns, rather than pairing each pretest raw score with its standard score, percentile score, etc., followed by each posttest score and its transformations. This greatly simplifies the mechanics of analysis; comparisons are nearly always made between pre- and posttest score of the same type.

School _____
Class _____
Year _____
Grade _____

## Cover Sheet for Data Recording Forms

| | Biodata | | | | Demo Data | Classi-fication | Treatment | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Student | Age in Sept | Yrs in U.S. | Lan-guage | Birth | | | Yrs in Pro-gram | Read-ing | ESL | Teach |
| | | | | | | | | | | |

School _____ Sheet _____ of _____

Class/Group _____ Recorder _____

Treatment/Comparison _____ Date _____

| Tests: | Pretest | Posttest |
|---|---|---|
| Name | | |
| Level | | |
| Form | | |

| Student Names | ID No. Cd | Grp | Sx | Ind | Date Pre | Post | Raw Score Pre | Post | Standard Score Pre | Post |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | | | | | | | | | | |
| 2. | | | | | | | | | | |
| 3. | | | | | | | | | | |
| 4. | | | | | | | | | | |
| 5. | | | | | | | | | | |
| 6. | | | | | | | | | | |
| 7. | | | | | | | | | | |
| 8. | | | | | | | | | | |
| 9. | | | | | | | | | | |
| 10. | | | | | | | | | | |
| 11. | | | | | | | | | | |
| 12. | | | | | | | | | | |
| 13. | | | | | | | | | | |
| 14. | | | | | | | | | | |
| 15. | | | | | | | | | | |
| 16. | | | | | | | | | | |
| 17. | | | | | | | | | | |
| 18. | | | | | | | | | | |
| 19. | | | | | | | | | | |
| 20. | | | | | | | | | | |
| 21. | | | | | | | | | | |
| 22. | | | | | | | | | | |
| 23. | | | | | | | | | | |
| 24. | | | | | | | | | | |
| 25. | | | | | | | | | | |

# ANALYZING THE DATA AND REPORTING THE RESULTS

MAJOR CONTENT ITEMS

8-A. DATA ANALYSIS CHECKLIST

8-B. REPORT-WRITING CHECKLIST FOR BILINGUAL-PROGRAM EVALUATORS

8-C. SAMPLE DATA-REPORTING TABLES

SECTION 8

## 8. ANALYZING THE DATA AND REPORTING THE RESULTS

Data analysis reporting is a complex undertaking that requires a person with adequate training. If such expertise is not available in the district, outside assistance should be sought. This section is written with the assumption that a competent evaluator will direct the analysis and focuses only on a few of the most common deficiencies encountered in educational evaluation reports.

A widespread problem in analyzing data is the failure to tie the analyses to the overall evaluation design. Many analyses simply do not answer the basic questions posed in the reports. Simple analyses that follow directly from the questions posed should be used. Sophisticated statistical approaches (e.g., multiple regression techniques), are usually not warranted and most smaller districts probably do not have the resources to employ such designs. Especially important, and widely ignored, is a careful examination of the data for obvious irregularities. Efforts in report writing should be focused on providing complete, but concise information rather than elaborate diagrams and exhaustive sets of uninterpreted data tables.

Key Problems

Grouping students for analysis. One of the major criticisms of bilingual program evaluations is that they lump together a wide range of students who have different characteristics and who receive a variety of (poorly described) services. Unless the reader of the report understands the characteristics of the students and the treatments they receive, discussions of achievement impacts will have little meaning. At an absolute minimum, students must be grouped for analysis according to language proficiency in both languages and according to the subjects they study (e.g., English reading, target language reading). If there are major differences in amount or type of instruction received by different students, then additional groups will be needed. (See Item 8-A.)

Presenting complete, convincing arguments. It is extremely rare to find an educational evaluation report that presents a complete argument for the existence of achievement impacts. Truly convincing reports are virtually unknown. Yet, presenting a reasonable argument is not difficult. The reader needs to know (a) the student characteristics, (b) the program goals, (c) the program features that are designed to achieve the goals, and (d) the results in terms of student scores. "Results," of course, must include the exact tests and procedures used. Finally, the relation between the treatment and the results must be summarized for the reader. These evaluation report basics are covered in Sections 2 through 8 of this manual and are summarized in Items 8-B and 8-C.

Related Issues

Floor and ceiling effects. Floor and ceiling effects are pervasive problems in bilingual-program evaluations. A minimal check, for multiple-choice tests is to be sure that mean classroom or school raw scores are no lower than 25 percent of the items correct for four-choice tests, 33 percent for three-choice, and so on. Mean raw scores should not exceed 75 percent of the total possible raw score on any test. Outside of these values, the likelihood of floor or ceiling effects, respectively, should be noted in the report.

Grade equivalent scores and other scales. Never use grade equivalent scores for any purpose. Use normalized standard scores (preferably NCEs) for all computations and calculations of impacts. Report pre- and post-test performance to general audiences in percentiles.

Statistical versus educational significance. Statistical significance says nothing about the size or importance of a program impact and should not be discussed in reports to general audiences. The real issue is whether the impact represents a noticeable reduction in the achievement problems to which the program is addressed.

Single-year versus longitudinal analysis. Most bilingual program evaluations are restricted to the effects of a single year. Such evaluations are not convincing. It is necessary to demonstrate that there is continuing year-to-year progress toward program goals.

Level of precision of the evaluation. Throughout this manual, the lack of precision of real-world educational evaluations has been emphasized, and the evaluation report should make this problem clear to the reader. On the other hand, if a program truly improves student achievement, this fact will show up clearly over a period of a few years in carefully conducted evaluations. Thus, while no single-year evaluation can be completely convincing, consistent results and trends over years will eliminate most doubts.

Executive and other summaries. The executive summary may be the most important part of the report since it will be the most widely read. The summary should cover all of the major report headings but should emphasize results and recommendations. (See Item 8-B.) Five to six pages should be enough. A copy of the executive summary should be written in the language (or languages) of the project parents and distributed to them.

## Data Analysis Checklist
Outline

I. General principles

A. Analyze data both by individual years for short-term goals and cumulatively for long term goals.

B. Separate data according to language proficiency groups.

C. Separate data further according to instructional treatment.

II. Preparation (applies to most evaluation designs)

A. Convert raw scores to standard scores (preferably normalized standard scores such as NCEs). Use these scores for all analyses.

B. Separate out those students with both pre- and posttests.
1. Compute means and standard deviations.
2. Plot the distributions of pretest scores.
3. Plot the distributions of posttest scores.
4. Plot the joint distribution of pretest and posttest scores.

C. For students with pretest scores only:
1. Compute the mean and standard deviation.
2. Plot the distribution of scores.

D. For students with posttest scores only.

Save the scores for student files and for use as next years pretest scores.

III. Check for irregularities in the data:

A. Floor or ceiling effects

B. Large changes in standard deviations from pretest to posttest.

C. Low correlations between pre- and posttest scores, or irregular joint distributions.

D. Differences between students who took the posttest, and those who dropped out.

E. Look for any other features of the data that strike you as strange, and be sure that you can explain them. Ideally, item data should be examined.

IV. Apply the statistical or other procedures relevant to the particular evaluation design in use.

Be sure that your analyses are relevant to the questions you are trying to answer.

## Report-Writing Checklist for Bilingual Program Evaluators

This checklist presents an outline that can be followed in preparing an evaluation report. The "Section Reference" to the right of each topic refers to the section of this manual that deals with the topic.

The purpose of the outline is to suggest one logical order of presentation of topics. There are, of course, other ways of organizing the report. A second function of the outline, however, is to provide a comprehensive reminder to the program director and evaluator of the kinds of information that may be included in a report. In the PIP field test evaluation reports, sections on student selection criteria and procedures, and interpretation of findings, were frequently not included. Such information should be included in a report to be considered complete.

## Report-Writing Checklist for Bilingual Education Program Evaluators

| | Section | Check when done |
|---|---|---|
| I. Executive Summary | 8 | ____ |
| A. Summary of findings | 8 | ____ |
| B. Recommendations | 8 | ____ |
| II. Program Overview and Background | 3, 8 | ____ |
| A. Brief program description | 3-A | ____ |
| B. Major goals | 2 | ____ |
| C. Context of program | 3-A | ____ |
| D. Program history and district needs | 3-A | ____ |
| E. Target student needs | 5 | ____ |
| III. Description of Evaluation | 4 | ____ |
| A. Purposes and audiences | 1, 8 | ____ |
| B. Evaluation staff and roles | 1 | ____ |
| C. Designs | 4 | ____ |
| 1. Questions addressed | 4 | ____ |
| 2. Comparison standards | 4 | ____ |
| 3. Constraints and questions not addressed | 4 | ____ |
| F. Continuity with previous and future years' evaluations | 4 | ____ |
| IV. Parent and Community Component | App. B-3 | ____ |
| A. Goals and objectives | App. B-3 | ____ |
| B. Description of activities | App. B-3 | ____ |
| C. Process evaluation | App. B-3 | ____ |
| 1. Measures used | App. B-3 | ____ |
| 2. Data collection procedures | App. B-3 | ____ |
| 3. Analyses and results | App. B-3 | ____ |
| 4. Interpretation | App. B-3 | ____ |
| 5. Recommendations | App. B-3 | ____ |

| | | | |
|---|---|---|---|
| D. | Outcome evaluation | App. B-3 | ____ |
| 1. | Measures used | App. B-3 | ____ |
| 2. | Data collection procedures | App. B-3 | ____ |
| 3. | Analyses and results | App. B-3 | ____ |
| 4. | Interpretation | App. B-3 | ____ |
| 5. | Recommendations | App. B-3 | ____ |
| V. | Staff Development Component | App. B-2 | ____ |
| A. | Goals and objectives | App. B-2 | ____ |
| B. | Description of activities | App. B-2 | ____ |
| C. | Process evaluation | App. B-2 | ____ |
| 1. | Measures used | App. B-2 | ____ |
| 2. | Data collection procedures | App. B-2 | ____ |
| 3. | Analyses and results | App. B-2 | ____ |
| 4. | Interpretation | App. B-2 | ____ |
| 5. | Recommendations | App. B-2 | ____ |
| D. | Outcome evaluation | App. B-2 | |
| 1. | Measures used | App. B-2 | ____ |
| 2. | Data collection procedures | App. B-2 | ____ |
| 3. | Analyses and results | App. B-2 | ____ |
| 4. | Interpretation | App. B-2 | ____ |
| 5. | Recommendations | App. B-2 | ____ |
| VI. | Students | 5 | ____ |
| A. | Selection criteria and procedures | 5 | ____ |
| 1. | Legal requirements | 5 | ____ |
| 2. | Make-up of program classrooms and definition of "project student" | 5 | ____ |
| 3. | Criteria for selection of students of limited English proficiency | 5 | ____ |
| a. | Tests and cutoff scores used | 5, 6 | ____ |
| b. | Role of teacher judgment | 5, 6 | ____ |
| c. | Role of parent wishes | 5, 6 | ____ |
| d. | Method of combining criteria | 5, 6 | ____ |

| | | |
|---|---|---|
| 4. Criteria for selection of students proficient in English | 5 | ____ |
| a. Criteria used | 5 | ____ |
| b. Method of application of criteria | 5 | ____ |
| 5. Exit criteria and follow-up | 5 | ____ |
| 6. Student turnover | 5, 9 | ____ |
| 7. Effects of selection criteria and procedures on evaluation design | 4, 5 | ____ |
| 8. Recommendations for improvement of entry/exist criteria and procedures | 5 | ____ |
| B. Description of students | 5 | ____ |
| 1. Characteristics at beginning of year | 5 | ____ |
| a. Language proficiency | 5, 6 | ____ |
| (1) English | 5, 6 | ____ |
| (2) Non-English language | 5, 6 | ____ |
| b. Achievement level | 5, 6 | ____ |
| c. Biographic data | 5-A | ____ |
| (1) County of birth | 5-A | ____ |
| (2) Years of residence in U.S. (if applicable) | 5-A | ____ |
| (3) Home language use | 5-A | ____ |
| (4) Previous educational experience | 5-A | ____ |
| (5) Other | 5-A | ____ |
| d. Demographic data | 5-A | ____ |
| (1) SES | 5-A | ____ |
| (2) Other | 5-A | ____ |
| 2. Current experience characteristics | | ____ |
| a. Attendance | 5-A, 9 | ____ |
| b. Key treatment variables | 3 | ____ |

| | | | |
|---|---|---|---|
| (1) | Reading in English | 3-C | ____ |
| (2) | Reading in non-English language | 3-C | ____ |
| (3) | Second language instruction | 3 | ____ |
| (4) | Participation in other special projects | 5-A | ____ |

| | | | |
|---|---|---|---|
| VII. | Instructional Component | | ____ |
| A. | Goals and objectives | 2 | ____ |
| 1. | Areas to cover | 2 | ____ |
| a. | Achievement | 2 | ____ |
| b. | Affect | 9 | ____ |
| 2. | Breakdown of goals and objectives by | 2 | ____ |
| a. | Grade level | 2 | ____ |
| b. | Language proficiency group | 2 | ____ |
| c. | Subject area | 2 | ____ |
| d. | Language of subject area | 2 | ____ |
| e. | Number of years of participation in project | 2 | ____ |
| 3. | Time frame | 2 | ____ |
| a. | Short-term goals | 2, 4 | ____ |
| b. | Long-term goals | 2, 4 | ____ |
| 4. | Explanation of bases for establishing criteria for success | 2 | ____ |
| 5. | Follow-up goals | 2 | ____ |
| B. | Description of instruction | 3 | ____ |
| 1. | Program-level instructional features | 3-A | ____ |
| 2. | Classroom-level instructional features | 3-B;3-D;9 | ____ |
| 3. | Reading instruction | 3-C | ____ |
| 4. | Level and extent of description | 3 | ____ |
| a. | Describe instruction at appropriate level (indiv. groups, classroom) depending on homogeniety of instruction | 3 | ____ |

| | | |
|---|---|---|
| b. Longitudinal description | 3 | ____ |
| 5. Characteristics of instructional staff | 3 | ____ |
| 6. Description of treatment received by comparison students or norming group | 3 | ____ |
| C. Process evaluation | | ____ |
| 1. Tests and measures used | 6 | ____ |
| 2. Data collection procedures | 7 | ____ |
| 3. Data analysis and results | 8 | ____ |
| 4. Interpretation of findings | 8 | ____ |
| 5. Summary of recommendationa made to improve instruction | 8 | ____ |
| D. Outcome evaluation | | ____ |
| 1. Tests and measures used | 6 | ____ |
| a. Relation of measures to goals | 2, 6 | ____ |
| b. Description of measures | 6 | ____ |
| (1) Language | 6 | ____ |
| (2) Content | 6 | ____ |
| (a) match between content of test and content of curriculum | 6-C | ____ |
| (b) cultural and linguistic appropriateness | 6 | ____ |
| (3) Technical properties | 6 | ____ |
| (a) validity and realibility | 6 | ____ |
| (b) floor and ceiling effects | 6 | ____ |
| (4) Form, level, edition | 6 | ____ |
| 2. Data collection procedures | 7-A | ____ |
| a. Explanation of which students were tested in which language(s) and rationale | 3, 5 | ____ |

| | | | |
|---|---|---|---|
| b. | Qualifications and training of testers, observers, interviewers | 7 | _____ |
| | (1) Training | 7 | _____ |
| | (2) Language skills | 7 | _____ |
| | (3) Familiarity with students, parents, etc. | 7 | _____ |
| c. | Schedules of data collection | 7 | _____ |
| d. | Scoring and recording | 7-B | _____ |
| 3. | Analysis and results | 8 | _____ |
| a. | Explanation of scales used | 6, 8, 8-A | _____ |
| b. | Floor and ceiling effects | 6, 8, 8-A | _____ |
| c. | Unit of analysis | 8, 8-A | _____ |
| | (1) By language proficiency group | 5, 8 | _____ |
| | (2) By treatment group | 3, 8 | _____ |
| d. | Explanation of irregularities | 7, 8 | _____ |
| | (1) Attrition | 8 | _____ |
| | (2) Bad data | 7, 8 | _____ |
| e. | Scope of analysis | 8 | _____ |
| | (1) Relation to previous years | 4, 8 | _____ |
| | (2) Plans for future continuity | 4, 8 | _____ |
| f. | Tables of test results | 8-C | _____ |
| 4. | Interpretation of findings in light of: | 8 | _____ |
| a. | Short-term and long-term goals | 2, 8 | _____ |
| b. | Degree of program implementation | 3, 8 | _____ |
| c. | Specific instructional treatment | 3, 8 | _____ |
| d. | Teacher characteristics | 3, 8 | _____ |

| | | |
|---|---|---|
| e. Similarities and differences between treatment group and comparison (or norm) group | 4, 8 | ____ |
| f. Number of years of student participation in project | 5, 8 | ____ |
| g. Match between tests and curriculum | 6, 8 | ____ |
| h. Limitations of tests | 6, 8 | ____ |
| i. Data collection procedures | 7, 8 | ____ |
| 5. Recommendations for improvement | | ____ |
| a. Instruction | 8 | ____ |
| b. Evaluation | 8 | ____ |

## SAMPLE DATA REPORTING TABLE

The use of data reporting tables enables the evaluator to provide a great amount of information in a concise and easy to read form. For the reader, tables provide an easy means of grasping the quantitative information a report has to offer.

In order to display data effectively, there are a number of information items that should be included. A table should identify what information is being provided, for what group or subgroup, and for how many participants (N).

In identifying the test used, the test edition year, form, language, and level should also be specified. It is also advantageous to report the number of items the test contains per subtest plus the date the test was administered. When reporting numerical data, it is necessary to identify pre- and posttest data, provide means, standard deviations, and gains.

A typical error is the failure to report the type of scores. The table should indicate whether scores are percentiles, standard scores, or raw scores.

Two sample data reporting forms are provided, one for reporting raw scores and the other for standard scores and percentiles. Raw score tables may be more useful to teachers who are familiar with the test. Percentiles and standard scores may be more useful for reporting to program administration and program monitors. The tables can be adapted to suit the needs of individual programs.

## SAMPLE DATA RECORDING TABLE

Program: _____

Language Classification of Group: _____

Grade: _____

Subject(s): _____

Test Description

| | Name | Language | Subtest (if used) | Level | Form | Edition | Norms Used (if any) | Testing Dates |
|---|---|---|---|---|---|---|---|---|
| Pretest | | | | | | | | |
| Posttest | | | | | | | | |

| Subtest(s) | N | Standard Scores (or NCEs) | | | | Percentile Equivalents | | Pre - Post Change |
|---|---|---|---|---|---|---|---|---|
| | | Pretest | | Posttest | | | | |
| | Avg. N | Mean | S.D. | Mean | S.D. | Pretest | Posttest | NCE Units |
| | | | | | | | | |
| | | | | | | | | |

Avg. = Average daily enrollment
N = The number of students who had both pretest and posttest

SAMPLE DATA RECORDING TABLE

Program: _____

Language Classification of Group: _____

Grade: _____

Subject(s): _____

Test Description

| | Name | Language | Subtest (if used) | Level | Form | Edition | Norms Used (if any) | Testing Dates |
|---|---|---|---|---|---|---|---|---|
| Pretest | | | | | | | | |
| Posttest | | | | | | | | |

Raw Scores

| Subtest(s) | No. of Items | N | Pretest | | Posttest | | Pre - Post Change |
|---|---|---|---|---|---|---|---|
| | | Avg. N | Mean | S.D. | Mean | S.D. | Raw Score Gains |
| | | | | | | | |
| | | | | | | | |

Avg. = Average daily enrollment
N = The number of students who had both pretest and posttest

# APPENDIX A

# HOW BIG ARE ACHIEVEMENT GAINS

## How Big Are Achievement Gains?

In order to be able to set realistic goals and to interpret gains made in bilingual programs, it is useful to have in mind the gains ordinarily made by English-speaking students in traditional all-English programs and in special programs such as Title I. The size of achievement gains resulting from special educational programs are generally small.

However, the differences between bilingual programs and traditional or other special all-English projects raise several issues that must be taken into account in considering the size of achievement gains. First, gains must be measured in the students' primary language as well as in English. This is important since much of the instructional time, at least in the early stages, is devoted to teaching content through the primary language and developing primary language skills. Second, since reading instruction in the second language may begin after reading skills are developed in the primary language, the grade level when English reading gains can be expected depends on the curriculum of the program. Third, it may be inappropriate to speak of gains relative to the norming population since English norms are not appropriate comparison standards for students of limited English proficiency, and no adequate norms for languages other than English are available at the time of printing (see Chapters 4 and 6).

### Normal Classroom Growth in All-English Traditional Programs

In order to discuss the size of achievement gains, it is necessary to have a meaningful standard or scale of measurement. For the purpose of this discussion, let us use the expanded standard score scale from a standardized test to provide a numerical score for general reading skill in English. A numerical score requires some frame of reference to give it meaning, and we can supply a useful frame of reference by identifying the ranges of reading scores for national norm-group students at various age levels. Figure 1 illustrates four norm-group distributions showing the range of English reading scores (10th percentile to 90th percentile) at the beginning and end of second grade and the beginning and end of

sixth grade. These percentile scales wil be used as our scales of measurement, and they can easily be converted to NCE units.[1]

Figure 1 illustrates the gain that 20th percentile norm-group students ordinarily achieve during a school year. (This percentile level was chosen for illustration since many Title I and Title VII students score in this range.) The white bar at the left of Figure 1 represents this amount, which we will refer to as "normal growth" for 20th percentile second graders. Also shown in Figure 1 is the amount of gain that constitutes "normal growth" for 20th percentile norm-group students at the sixth-grade level. It can easily be seen that the sizes of these gains vary across grade levels, a point that will be further discussed below.

We are now in a position to compare student gains with percentile levels. For example, it can be seen from Figure 1 that a Title I student who started the second grade at the 20th percentile would have to gain nearly <u>twice</u> as many points as the 20th-percentile students with normal growth in order to reach the 50th percentile in the spring. A 20th-percentile sixth grader would have to achieve over <u>four times</u> normal growth to reach the 50th percentile by spring.

Moreover, it is important not to conclude even for the second grade that doubling the amount of instruction or doubling the effectiveness of the instruction would be enough to raise the student to the 50th percentile. Normal growth is certainly due in part to classroom instruction, but it also includes all the effects of out-of-school learning and maturation, and these effects cannot be doubled so easily. It is also affected by the motivation of the student. In other words normal growth is a result of:

- classroom instruction,
- out-of-school learning,
- maturation,
- motivation.

---

[1]NCEs are normalized standard scores with a mean of 50 and a standard deviation fo 21.06. Because the scale is normalized it is assumed to be equal-interval--that is, the length of the interval between any two adjacent scores in the scale is equal to the interval between every other pair of scores (Tallmadge and Wood, 1976).

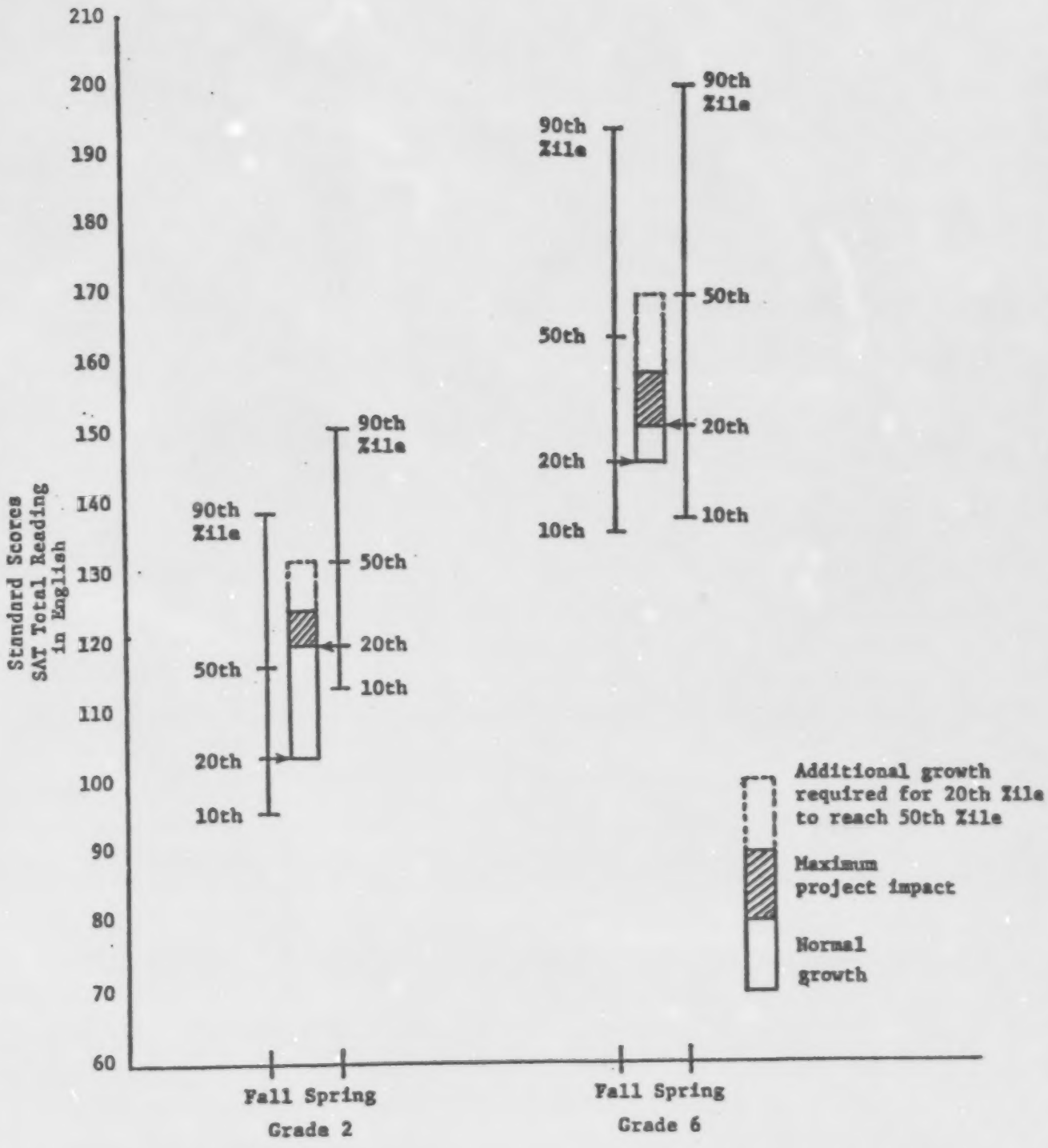



Figure 1. Impacts of regular classroom learning and Title I projects on achievement test scores in English.

183

Thus, even if we can double the amount or the effectiveness of classroom instruction, we should not expect to double the amount a student learns.

Impact of Title I projects. If it is true that the classroom is only one of several factors contributing to student learning, then even dramatic improvements in school instruction might produce rather modest gains. Existing data, though not conclusive, tend to bear this out. Analysis of data from a great many exemplary Title I projects suggests that, in terms of the scale in Figure 1, gains produced by projects are small. In fact, it is difficult to find convincing evidence of gains of even one-third of a standard deviation with respect to the national norm.[2] This amount has been added to the bar in Figure 1 to represent the maximum impact that might be expected from an exemplary Title I project. Of course, this is not a rigorously established limit, but based on available program evaluations, it appears to be a realistic value.

To complete the picture, consider the growth scales in Figure 1 for sixth graders. Note that the spread between the 10th and 90th percentiles is greater for the older age group, but that normal growth for 20th-percentile students is considerably less than at second grade. This normal growth still includes the effects of out-of-school learning, maturation, and motivation, so the maximum Title I impact (again represented in the figure as one-third of a standard deviation) would require a project that was more than twice as effective as regular classroom instruction alone.

In short, normal growth looks rather small when measured against the percentile scale, and the amount of growth that can be directly attributed to classroom instruction is even smaller. Thus, even a dramatically effective Title I program, one in which instruction is several times as effective as that in the regular classrom, may raise student scores by only

---

[2]This amount has been suggested as representing a "just noticeable difference" when comparing two groups on physical attributes such as height or weight (A. O. H. Roberts, 1977b). In the context of project evaluation, it has been used as an arbitrary criterion below which gains were considered of little educational significance (Tallmadge and Horst, 1976).

a few percentile points or NCEs per year. In most cases, evaluations have been designed to measure much larger gains than we can reasonably expect to find. In such evaluations, the relatively small program impacts that actually occur may be completely obscured by the amounts of error normally associated with an evaluation.

Impact of Bilingual Education Projects

In bilingual education programs it is more difficult to make generalizations about the size of achievement gains for a number of reasons. In order to discuss the size of gains it is necessary to have a meaningful scale of measurement (scores that can be referred to a familiar range of scores). Unfortunately, such scales of measurement are available only for major achievement tests in English, although work is currently being done to develop meaningful scales for some language proficiency tests and Spanish achievement tests.

It is also necessary to have meaningful standards of comparison in order to determine whether the amount of growth, as measured on the scale, is greater or less than the amount of growth that would be made by similar students participating in (1) similar bilingual programs, or in (2) traditional, all-English classrooms. Some tests have norms designed to provide a standard of "normal growth" in bilingual programs, but such norms are not yet considered technically adequate. Standards that provide an adequate no-treatment expectation for students of limited English proficiency simply are not available. (See Section 6, Limits to the Usefulness of Norms). Some project personnel have asked why they cannot simply use the English norms for the Spanish version of a test. This would be highly inappropirate since it cannot be assumed that students in a bilingual program grow at the same rate as students in traditional norming populations (either in English or in their native language). There is no evidence that the equipercentile assumption holds true for students of limited English proficiency. In other words, it cannot be assumed that a group of limited-English students who score at the 30th percentile, for example, at pretest time would score at the 30th percentile at posttest time. They

may exceed "normal growth" in some subjects at certain times and fall below it at other times depending on a complex of factors.

In bilingual education programs the size of the gains that can be expected, the language in which gains can be expected, and the grade level at which they can be expected depend on a number of factors. These include:

1. the type of students of limited English proficiency being served,
2. the program model being implemented, and
3. the general context in which the program operates.

For example, consider the case of students who are relatively balanced bilinguals from the time they enter school, but are also somewhat limited in English proficiency, and are receiving English instruction in all subject areas as well as some native language instruction. There are some reasons to expect that these students may make gains in English achievement similar to those made by students in Title I programs. If these students receive instruction primarily in the dominant language, English gains for that year may be less than the norms and larger gains will be expected to appear after transfer to a greater amount of instruction in English. Now consider the case of students who are extremely limited in English proficiency at the time of pretesting. It seems reasonable to assume that achievement as measured by English tests may exceed "normal growth" if a great deal of English instruction is provided. If instruction is provided primarily in the dominant language, English gains may be less than "normal growth" and will be expected to appear at a higher grade level when transfer occurs.

Gains that can normally be expected will be discussed in relation to the three subject areas most stressed in bilingual programs (a) oral language, (b) reading, and (c) math. Under each topic gains will be discussed for the native language and for English.

## Oral Language

Although oral language development in both the first and the second language is a major goal of bilingual programs, there is little data

available to indicate what kind of growth can be expected in bilingual programs. Many programs administer oral proficiency tests for purposes of classification, but unfortunately gains are often not reported as part of the impact evaluation. No major studies have examined this issue, probably due to the time involved in individual test administration.

Native language. The amount of growth that can be expected in the native language depends on students' initial proficiency in the home language, the amount and quality of instruction, and language use in the home and community. Although norms are available for some tests, there is little data to indicate what "normal growth" is. In the bilingual PIP field test, only 7 of 19 sites reported both pre- and posttest data for oral proficiency in the native language. Due to the variety of tests used and the different types of scores reported (language proficiency level vs. raw score), it is not possible to arrive at any generalizations. Nevertheless, one important point should be made. Two sites demonstrated a loss in Spanish language proficiency on the Language Assessment Scales. While for English language tests it is assumed that a certain amount of growth from fall to spring is inevitable, due to maturation and other influences, this is not necessarily true for non-English languages in the U.S. Native language loss may be the norm in certain programs in certain communities.

English. There is also a lack of information concerning "normal growth" in English oral language. Logically, it might be expected that large gains would occur during the first and second years of a student's participation in a program since curricular emphasis is placed on second language acquisition. Data from the PIP field test indicate that on one commonly used test, the Language Assessment Scales, students (n = 162) in the first grade gained an average of 13.7 raw score points per year (in 7 sites) and students (n = 102) in the second grade (in 3 sites) gained an average of 8.7 points (out of 100 total points possible).

Reading

Native language. The amount of growth that can be expected in native language reading scores depends on initial native language proficiency of the students, the instruction provided, and the degree of use of the language in the home and community. Norms are available for some Spanish language achievement tests (see Section 6). Although these norms cannot provide a no-treatment expectation, and are not completely adequate technically, they can provide a very rough estimate of reading growth expected.

English. Some data on English reading gains are available from a national study of Title VII bilingual programs.[3] Although the study has some serious methodological flaws, such as a comparison group that was not sufficiently comparable to the treatment group, there is some information in the report that can be useful to evaluators. Students in a national sample of bilingual programs were pre- and posttested with the CTBS Reading test over a period of about five months. Table 1 illustrates average amounts of growth for students. The pre- and posttest scores are expressed in percentiles, and the amount of pre to post growth is expressed in percentiles and NCEs. The scores are reported for grades 2 through 6 for four different groups of students who were classified into language proficiency groups by their teachers for purposes of test taking. For example, an "English-dominant bilingual" was defined as a student whose teacher felt that s/he should take a reading test in both languages and a math test in English. The percentile scores represent the students' standing relative to a national norming sample. If there is no change in percentile standing from pre to posttest, then it is usually assumed that "normal growth" has occurred for the students at that level. We do not know, however, how similar students would have performed without a bilingual program. These norms do not provide a no-treatment expectation; they only serve to compare

---

[3] American Institutes for Research. Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Volume I: Study Design and Interim Findings, February, 1977.

students to the national average. A gain of 7 or more NCEs is considered large for purposes of this discussion. A gain of 3 to 6 NCEs is moderate and a gain of 2 or less is minimal.

Monolingual Spanish speakers scored from the 2nd to the 5th percentile on pretests and from the 3rd to the 6th percentile on the posttests. They showed gains ranging from 0 to 1 percentile points (0 to 3 NCEs). Lack of substantial growth in this area might be attributable to several factors, among them (1) floor effects that limit the extent to which gains can be detected, (2) curricular emphasis on English oral language development, and (3) reading instruction in the native language prior to introduction of English reading.

For the Spanish-dominant bilingual group, there were very large gains of 13 percentile points (16 NCEs) in the second grade. This may be due to increased English language proficiency as well as improved reading ability. In many programs second grade students have had very little, if any, experience in English reading at pretest time, but by the end of the year they have transferred reading skills to English. Gains were moderate in the 5th grade, minimal in 4th and 6th, and there is a moderate loss in the 3rd grade relative to the national norms.

English-dominant bilinguals demonstrated moderate gains in the 2nd and 5th grades, and "normal growth" relative to national norms in the 3rd, 4th, and 6th grades. The monolingual English group showed a gain of 12 percentile points (9 NCEs) at second grade, moderate losses at the 3rd and 5th grades, and minimal changes in percentile standing in the 4th and 6th grades. It should be noted that these second graders scored at the 15th percentile level at pretest time, while the English dominant bilinguals started out substantially higher at the 25th percentile.

Caution must be exercised in interpreting "losses" relative to national norms. A drop in percentile standing (for example from the 18th to the 16th percentile) could represent positive program impact if similar students would have dropped even further without a bilingual program.

Mathematics

Table 1 also displays percentile scores and gains made by the average Title VII student in mathematics computation. Monolingual English students and English-dominant bilinguals took the test in English, while the other two groups took the test in Spanish. The pattern is quite different from that of the reading scores. Of the 20 groups reported across grade levels and language groups, 17 made gains relative to the national norms.

The monolingual Spanish group in the 2nd and 3rd grades made gains of 16 and 13 percentile points respectively (10 and 9 NCEs) relative to national norms, while moderate gains were demonstrated in the 4th, 5th, and 6th grades. The Spanish-dominant bilingual group showed very large gains at second grade (30 percentile points, 17 NCEs), moderate gains at 3rd and 4th grade, minimal change at 5th grade, and a moderate loss at 6th grade.

The English-dominant bilingual group started with the highest 2nd-grade percentile standing of the four language dominance groups at pretest time: 37th percentile compared to 28th for the monolingual English group and 17th and 18th for the other two groups. The English-dominant bilinguals exhibited moderate gains at the 2nd, 4th, and 6th grades, a minimal gain at 3rd grade and "normal growth" at the 5th grade. The monolingual English group exhibited large gains at the 3rd grade level (11 percentile points, 7 NCEs), moderate gains at 2nd and 4th, a minimal gain at 5th, and a moderate loss at 6th.

In summary, these data indicate that growth patterns for students in bilingual programs differ from growth patterns of students in a national norming sample. For example, very large gains in national percentile standing were demonstrated in English reading and math for Spanish-dominant bilingual students in the second grade. The information presented here may provide some assistance to districts in setting goals. A word of caution is in order, however. These data represent bilingual programs of a variety of types and it is not possible to determine which type of program

produced which gains. There may be interactions among program type, language group, and grade level that affect these data. In addition, there may also be large amounts of error due to floor effects for some groups. Gains to be made in individual projects will depend on a number of factors including characteristics of students and the type and quality of instruction provided.

Table 1
National Percentiles (NP) for CTBS Reading Total and Mathematics Computation Means by Judged Language Dominance Group--Title VII Hispanic Students[1]

| Outcome Variable | Grade | Monolingual English | | Change | | English-Dominant Bilingual | | Change | |
|---|---|---|---|---|---|---|---|---|---|
| | | Pretest NP | Posttest NP | %ile | NCEs | Pretest NP | Posttest NP | %ile | NCEs |
| CTBS Reading Total Score (English) | 2 | 15 | 27 | 12 | 9 | 25 | 29 | 4 | 3 |
| | 3 | 25 | 21 | -4 | -3 | 25 | 24 | -1 | -1 |
| | 4 | 30 | 28 | -2 | -1 | 22 | 24 | 2 | 1 |
| | 5 | 22 | 18 | -4 | -3 | 19 | 27 | 8 | 5 |
| | 6 | 20 | 21 | 1 | 1 | 18 | 18 | 0 | 0 |
| CTBS Mathematics Computation Score (English) | 2 | 28 | 38 | 10 | 6 | 37 | 46 | 9 | 5 |
| | 3 | 29 | 40 | 11 | 7 | 32 | 36 | 4 | 2 |
| | 4 | 37 | 48 | 11 | 6 | 42 | 48 | 6 | 3 |
| | 5 | 32 | 33 | 1 | 1 | 33 | 33 | 0 | 0 |
| | 6 | 40 | 34 | -6 | -4 | 34 | 38 | 4 | 3 |
| | | Spanish-Dominant Bilingual | | | | Monolingual Spanish[2] | | | |
| CTBS Reading Total Score (English) | 2 | 5 | 18 | 13 | 16 | 5 | 6 | 1 | 2 |
| | 3 | 17 | 13 | -4 | -4 | 4 | 4 | 0 | 0 |
| | 4 | 14 | 15 | 1 | 1 | 3 | 3 | 0 | 0 |
| | 5 | 18 | 25 | 7 | 5 | 3 | 3 | 0 | 0 |
| | 6 | 33 | 36 | 3 | 1 | 2 | 3 | 1 | 3 |
| CTBS Mathematics Computation Score (Spanish) | 2 | 19 | 49 | 30 | 17 | 18 | 34 | 16 | 10 |
| | 3 | 28 | 39 | 11 | 6 | 20 | 33 | 13 | 9 |
| | 4 | 49 | 58 | 9 | 5 | 31 | 41 | 10 | 5 |
| | 5 | 30 | 31 | 1 | 1 | 19 | 26 | 7 | 4 |
| | 6 | 45 | 38 | -7 | -3 | 23 | 31 | 8 | 6 |

[1]These data are taken from: American Institutes for Research. Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Volume I: Study Design and Interim Findings, February, 1977.

[2]Monolingual Spanish students at all grade levels took the CTBS/S, Level C, as both a pretest and posttest. CTBS norms used here were those for the form and level taken by all other judged language dominance groups at a particular grade level.

# APPENDIX B

# GUIDELINES FOR OTHER EVALUATION AREAS

Contents

B-1. Evaluation of Affective Impacts

B-2. Evaluation of Staff Development

B-3. Evaluation of Parent/Community Involvement

## Introduction

These sections address the evaluation of student affective growth, staff development, and parent/community involvement. Virtually all bilingual education programs have goals in these areas and spend considerable time and funds in activities designed to address them. There is a lack of information, however, on methods and issues in evaluating these areas, and program evaluations, often due to time constraints, often give them low priority.

In an attempt to improve local evaluations and to broaden their scope, RMC provided technical assistance to the bilingual PIP field-test sites. The evaluators were encouraged to (a) employ measures other than achievement tests, and (b) to evaluate goals other than student achievement goals.

The following procedure was used in developing these sections: (a) evaluation reports from the PIP sites were reviewed and analyzed; (b) current relevant literature was reviewed; (c) recommendations and suggestions were developed; and (d) materials were sent to all 19 PIP field-test sites in the hope of improving the evaluations.

The format of Appendix B differs from that of the previus eight sections since each Section (1,2,3) consists of the document that was sent to each site participating in the bilingual-PIP field test. Each section contains the following: (a) a review of practices employed by bilingual programs to evaluate the component, including the most common practices and other practices used by at least one site; (b) a discussion of technical issues; and (c) recommendations for improving the evaluation of this component.

In addition, the section on evaluation of affective impacts includes recommendations for the use of two unobtrusive measures of project impact: attendance and retentions.

# EVALUATION OF AFFECTIVE IMPACTS

## Common Practices

Most bilingual projects have explicit goals for student affective growth. The most common goals of the projects in the Bilingual PIP field test study were to:

- increase awareness of and appreciation for the child's own culture and the dominant culture, and to
- improve self-concept.

A number of sites that had stated affective goals employed no measures and reported no results in this area. Of those sites that did address student affect, the most common approaches were the following:

- paper-and-pencil, self-report measures of self-concept, administered pre and post;
- paper-and-pencil, self-report measures of cultural attitudes, administered pre and post;
- documentation of classroom and outside cultural activities offered by project;
- reporting the percentage of students who paticipated in a given number of cultural events in the classroom.

Other approaches used by at least one site were:

- teacher rating scale to assess students' social behavior;
- teacher rating scale to assess students' school-related behavior and attitudes;
- teacher rating scale to assess student attitude toward self as a bilingual and toward others as bilinguals;
- teacher rating scale to assess students' participation in classroom and playground;
- paper-and-pencil, self-report measures of attitude toward school and toward school subjects, administered 3-4 times during year.

## Pro edures Recommended to Tryout-Sites

Immediate effects on students. In attempting to describe the effects of a bilingual project on the LES students, it is essential to examine them from a broader perspective than simply noting changes that occur over one

year. Evaluators should describe both immediate and cumulative effects. The very nature of a bilingual project makes it different from other types of special projects in one important way. In most special projects, it is assumed that the normal treatment is meaningful to the students, at least in the sense that that they can comprehend the language used in instruction, but that the special project consists of a better method of teaching. The situation is different in a bilingual program. The normal, all-English program cannot be "meaningful" (in the sense of the Lau decision) if children are not yet fluent speakers of English. Instruction is meaningful to children only to the extent that they can understand what is said to them and participate in verbal exchanges with teachers and other students throughout the day.

For this reason, the first question that needs to be addressed by districts in evaluating effects of the project on students is: To what extent are students receiving a meaningful education? This question can be broken down into other questions such as: To what extent can teachers and children communicate with one another? What proportion of the day is meaningful to children in terms of the degree to which they speak and comprehend the language of instruction? To what extent are children able to relate to and profit from the instructional materials? These are complex questions to answer due to the range of language proficiency levels of children and the inadequacy of measurement techniques; nevertheless these immediate benefits to children should be addressed, since, although they are obvious to bilingual educators, they are not always obvious to others, and since, although the long-range effects of such instruction should show up in test scores, this is not always the case due to short-range evaluation designs and poor tests administered under questionable conditions.

Specification of goals. Measuring benefits to students in the affective domain is a tricky business for a number of reasons. The goals for the affective domain are often broad and vague. For example, the goal of improving self-concept is open to many interpretations. It is a controversial goal as well since it is not clear that LES students necessarily have low self-concepts, nor is it clear what the causal relationship is between self-concept and achievement. It might be made more specific and more manageable by breaking it down into various components. A project might set a goal that students in the bilingual project will improve their opinion of themselves as successful readers, for example, or as successful math students.

Causes of affective changes. Secondly, it is not clearly stated in most proposals and evaluation reports precisely why project features are expected to bring about changes in student attitudes. In some projects it is expected that self-concept will improve through an understanding of the cultural heritage associated with both languages (see, for example, Venceremos Project Management Directory, p. 86). For others it is implied that improved attitudes toward self and others are expected as a result of (1) accepting and using the language of the child; (2) providing successful learning experiences; (3) integrating the culture of the child into the curriculum; (4) involving parents in classroom and other activities; and

(5) employing bilingual, bicultural teachers who serve as role models. For still other sites it is implied that the project as a whole will bring about affective changes in students.

Many projects measure one chosen aspect of student attitudes and report the results without providing a discussion of the possible reasons for the results. If improvements are expected to be due to one of the project features mentioned above, then a crucial step must be to state whether that particular feature was implemented. For example, if project personnel expect the cultural component to influence student self-concept, then it would be useful to describe the nature and extent of the cultural component that was actually implemented. If there was no cultural component implemented, or if it was very minimal, then there is no reason to expect that it (or a lack of it) affected self-concept. Likewise, if improved self-concept is expected to be a result of the introduction of concepts in the native language, and the latter did not occur, there is no reason to expect to achieve the affective objective. Evaluators should state, to the extent possible, which project features, or combination of features, are expected to produce affective changes. They should then discuss to what degree those features were implemented. If they were not implemented, or were improperly implemented then it is not possible to attribute changes in student affective characteristics to those features. It is suggested that evaluators focus on the processes that are expected to bring about changes to see that these processes are in fact occurring.

Measurement. Bilingual PIP field-test sites that used affective measures made an attempt to locate the best measures available, but the choice of adequate measures (particularly in two languages) is very limited. Most sites used paper-and-pencil, self-report instruments or teacher rating scales. Self-report instruments are very unreliable for young children since social desirability and events of the moment have a great influence on responses. Teacher rating scales are more likely to be reliable, particularly if several measures are taken longitudinally. A variety of unobtrusive measures can also be used. Although there are always serious questions of validity and reliability associated with any affective measure, sites should choose the best measures possible.

One site administered an affective test and did not report results claiming that the test was not valid and reliable. Another site reported results of a locally developed measure, but discounted the results for similar reasons. If the reliability and validity of a locally developed test are unknown, these parameters should be investigated. If this is not possible, then it might be better to choose a commercially available instrument with established psychometric properties.

Evaluating affective changes is problematic since it is impossible to measure attitudes directly. Since an attitude is a hypothetical construct generally considered to be composed of feelings, behaviors, and knowledge or beliefs, it is necessary to choose possible indicators of an attitude, measure these, and make inferences about the attitude. Some suggestions concerning the kinds of attitudes that can be measured and

the possible manifestations of these attitudes are presented in the outline entitled "Approaches to Evaluating Affective Impacts." The outline includes approaches to (1) evaluation of the immediate effects on students, (2) evaluation of the instructional strategies intended to bring about attitudinal changes, and finally (3) evaluation of the attitudinal changes. Each item preceded by a bullit (o) is simply an example and there may be many others. The purpose of this report and the outline is to assist sites in exploring the variety of ways in which a district can describe project benefits to students. The number of approaches used and the extent of their use will depend, of course, on time and financial constraints. It is hoped that the suggestions provided here will assist districts in making better informed choices based on a number of options.

## Approaches to Evaluating Affective Impacts

### A. Evaluation of Immediate Project Benefits to Students

1. Instructional features contributing to a "meaningful" education (in the sense of Lau)

   - presence of teachers and teacher aides who speak language of child
   - acceptance of and use of language of child for instructional and other purposes
   - use of instructional materials written in language of child

2. Immediate potential effects on students

   - ability to communicate with teachers and other students
   - ability to profit from instruction and participate more fully in other activities
   - ability to relate to and profit from instructional materials

3. Measurement techniques

   - language-use observation instrument
   - teacher self-report of language use
   - teacher rating scale or questionnaire to evaluate materials

### B. Evaluation of Processes Leading to Affective Changes in Students

1. Processes expected to lead to affective changes in students

   - integrating child's culture into the curriculum
   - providing successful learning experiences
   - accepting and using the language of the child
   - establishing good relations between home and school
   - providing role models
   - teaching the minority language to the majority group

2. Measurement techniques

   - classroom observation
   - interviews with appropriate staff
   - rating scales
   - self-report in upper grades

C. Evaluation of Student Attitudes

1. Attitudes toward self

   a. kinds of attitudes toward self
      - successful reader
      - in control (locus of control)
      - bilingual
      - successful math student
      - motivated
      - active participant in classroom
      - ethnic group member
      - creative and able to contribute

   b. manifestations of attitudes toward self
      - student comments
      - student non-verbal behavior
      - student language use

   c. measurement techniques
      - teacher rating scale
      - self-report, paper-and-pencil test
      - student interview

2. Attitudes toward others

   a. components of attitudes toward others
      - respect for other races or ethnic groups
      - respect for other cultures
      - respect for other languages

   b. manifestations of attitudes toward others
      - language use at school
      - interethnic play at school

   c. measurement techniques
      - sociometrics
      - language-use observation instrument
      - interethnic interaction observation instrument
      - rating scale

3. Attitudes toward school

   a. kinds of attitudes toward school
      - sense of belonging
      - particular school subjects (e.g., attitude toward reading)
      - school subjects in a particular language (e.g., Spanish reading)
      - academic and social activities

b. manifestations of attitudes toward school
   - attendance
   - active participation in activities
   - student comments
   - retentions
   - willingness to share school experiences with family

c. measurement techniques
   - attendance records
   - teacher rating scale
   - observation instrument
   - self-report paper-and-pencil test
   - retention records
   - parent interview or questionnaire

4. Attitudes toward home

   a. components of attitudes toward home
      - home language
      - home culture
      - alienation between home and school

   b. manifestations of attitudes toward home
      - willingness to speak home language even after English is mastered
      - willingness to share items and stories from home

   c. measurement techniques
      - rating scale
      - classroom observation
      - self report

## Unobtrusive Measures of Project Impact: Attendance and Retentions

Project evaluations generally rely heavily on tests and questionnaires. Since these techniques require the cooperation of a respondent, a great deal of time is often involved, and the measure itself can contaminate the response. Teachers already complain of overtesting, and it is difficult to obtain valid and reliable test results for children in the early grades.

For these reasons, we encourage sites to broaden their range of evaluation methodologies and to consider exploiting a variety of measurement possibilities. No measurement technique is without bias, but combining measurement techniques with different kinds of biases can give a more complete picture of what has occurred during the life of a project.

Unobtrusive measures do not require any sort of response and so do not interfere in any way with the students' school day. Two such measures worth considering for use in your district are records of attendance and retentions. If there is any indication that attendance has improved or retentions have been reduced as a result of the project over the last two years (or longer, if a bilingual project was already in operation), then these data would be worth examining. It is particularly important to examine these issues if they were addressed in a needs assessment or are project goals.

The chart illustrates comparisons which might be employed. Retentions in the project school for the current year (A) can be compared to retentions before the project was installed (B). The current project school (A) can be compared to a similar school that has no well established bilingual project (C). If neither of these comparisons is possible, then district, regional or national historical data can be used (D).

No matter which comparison is chosen (A to B, A to C, or A to D), be sure to compare the project group to a comparison group with similar characteristics. For example, one might make any or all of the following comparisons:

| | | | |
|---|---|---|---|
| 1. | LES | to | LES |
| 2. | Spanish Surnamed | to | Spanish Surnamed |
| 3. | FES | to | FES |
| 4. | Total School | to | Total School |

It is most desirable to compare LES project students to LES comparison students. If, however, the language proficiency characteristics of comparison students are not known, or if the criteria for designating comparison students as LES were substantially different from criteria being applied for project students, then it would be better to compare all Spanish-surnamed project students to all Spanish-surnamed comparison students. Gathering data on FES students or on the total school population serves the purpose of establishing the comparability of the project school(s) to the comparison school(s).

Interpretation of results must be made in light of the comparison used and must take into account the limitations of the available data. While there may be other influences which affected retention patterns, such as major policy changes, if it is likely that results are due at least in part to the project, they should be reported.

We have included two worksheets, one for gathering attendance data and one for retention data. They may be of assistance to you should you decide that this information would be appropriate for your evaluation. We would appreciate your comments on these worksheets. If you are already using a similar procedure or have relevant information we might share with other sites, please let us know.

## Possible Comparisons for Attendance and Retention Data

A. Current project school, 1978-79

B. Same school before project, 197_

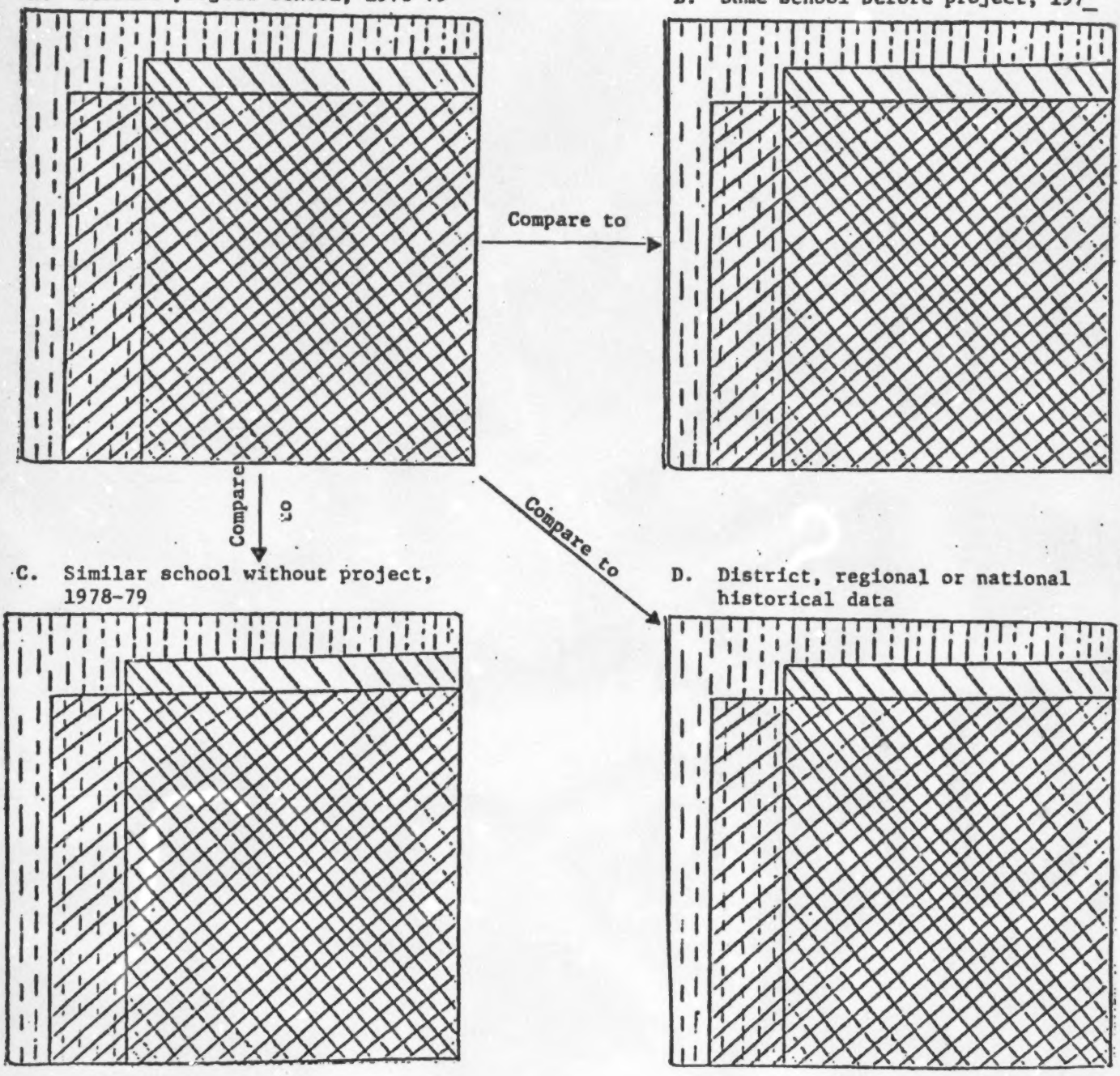
Compare to

Compare to

Compare to

C. Similar school without project, 1978-79

D. District, regional or national historical data



Key:

FES

NES/LES

Spanish-surnamed

## Retentions Worksheet

**Steps**

1. Determine what data is available.
2. Choose a comparison.
3. Calculate retentions of appropriate groups listed below for project students.
4. Calculate retentions of the same groups for comparison students.

**LES Retainees** (What percent of LES project students were retained?)

a. Number of LES project students retained in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

b. Total number of LES students in project in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

c. Percent of LES project students retained (a/b = %):
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

**Spanish-surnamed Retainees** (What percent of Spanish-surnamed project students were retained?)

a. Number of Spanish-surnamed project students retained in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

b. Total number of Spanish-surnamed students in project in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

c. Percent of Spanish-surnamed project students retained (a/b = %):
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

**FES Retainees** (What percent of FES project students were retained?)

a. Number of FES project students retained in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

b. Total number of FES students in project in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

c. Percent of FES project students retained in 1978-79 (a/b = %):
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

**Total School Retainees** (What percent of students in the project school(s) were retained?)

a. Number of retainees in project school(s) in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

b. Total number of students in project school(s) in 1978-79:
K _____ 1st _____ 2nd _____ 3rd _____ 4th _____

## Attendance

### Project school

Source: Each teacher's attendance records.

Method: Calculate mean percent of absences for students in project school.

Example:
1. For each student, divide the number of days absent by the number of days enrolled. (Example: 11/176 = 16%)
2. Calculate the mean percent of absences for the entire group.

### Comparison school

Source: Each teacher's attendance records.

Method: Same as for project students.

### Alternative methods of calculating attendance

Methods:

- Calculate mean attendance for only those students who were enrolled for at least 85% of the school year (which is an estimate of a student's enrollment period between pre- and posttesting.
- Calculate mean attendance for those students who participated in pre- and posttesting.
- A less desirable approach but perhaps more realistic in order to have an accurate comparison is to calculate mean attendance for all project students enrolled during the year. This calculation would include students who may have been enrolled for any period of time and moved away. This approach can be used when the comparison group does not have data which exclude the mobility factor. The comparison can be either a non-project school or historical data available from a particular school or, as a last resort, district historical data (see chart).

Possible Interpretations of Positive Results

Parents: Parents may see the project as more relevant and beneficial, therefore they may be more persuasive in seeing that their children attend school.

Student: Students may find their school experience more relevant and less traumatic, and therefore they may be more inclined to attend school.

School: The school may be affecting the behavior of students and parents by providing better school-home relations, thereby positively affecting students' attendance rate.

Attendance Reporting Form

| | 1 NES/LES | | 2 Spanish Surname | | 3 FES | |
|---|---|---|---|---|---|---|
| Grade | Project | Comparison | Project | Comparison | Project | Comparison |
| K | | | | | | |
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |

*Use the sections appropriate for your site depending on what information may be available and what questions your project wants to answer. For example, if your district does not have attendance records broken down by Language group, you may wish to use column 2 only.

# EVALUATION OF STAFF DEVELOPMENT

## Common Practices

A review of evaluation reports from sites participating in the field test of Bilingual Project Information Packages revealed that the most common approach to the evaluation of the staff development component was to:

- Provide description and/or documentation of workshops and other training activities that were provided, and to
- Evaluate the content of the training activities.

The description of workshops and other activities usually consisted of a list and some sample outlines of presentations. In order to evaluate the content of the sessions, most sites had workshop participants fill out a combination rating form/questionnaire in which they evaluated sessions in terms of criteria such as expertise of presentor, relevancy, clarity, practicality, meeting stated objectives, and meeting needs. The results of these evaluations were summarized across participants and often actual comments made by participants were included in the summary. Several such summary sheets, representing several workshops, were generally included in an appendix. The results were then summarized across several or all sessions for the year and the conclusion reached was often something like "With one exception, all workshops met their objectives and provided useful practical information for teachers." The majority of sites evaluated their staff development component at this level.

A number of sites employed additional techniques including the following:

- A needs assessment administered in the fall.
- Pre- and posttests on content of workshop administered to participants at each workshop.

- Classroom observation to determine areas in which training is needed.

- A questionnaire administered to a sample of district (non-project) staff to determine the extent to which they received information concerning the project.

- A pre-post (fall-spring) test for project staff measuring knowledge of cultures represented in class.

- A questionnaire to assess knowledge (self report) of project's goals and objectives.

- Reporting of university credits, special certificates, or degrees received during the project life.

These approaches from a sample of programs represent an attempt to conduct a broader evaluation of the staff development component. Since staff development is the main approach to implementing many bilingual programs, it is important to select evaluation strategies that will provide as thorough and accurate an assessment as possible of the effects of training.

Recommendations

The staff development component can be evaluated through a variety of approaches depending on who or what is evaluated, in terms of what specific qualities or characteristics, and over how much time. What is evaluated has generally been limited to the content of the pre- or in-service sessions. But to adequately assess the value of a staff development program, the effects of training sessions on program staff must also be examined. One general goal is to improve the teachers' and aides' performance in bilingual instruction, and the results can be determined by answering the following questions: How has classroom performance changed?, How have knowledge, skills, and attitudes changed?, How have language skills improved? An adequate evaluation should try to answer these questions. Another goal that has been receiving increased emphasis is the

upgrading of management and evaluation skills for program staff who perform these functions. It is just as important for project directors to be trained in communication skill, for example, as it is for teachers to be trained in instructional techniques.

The ultimate benefits of a staff development component should be its effects on the quality of the students' education. It is more difficult to measure effects on students and to be able to attribute them to training sessions, but, in some cases, districts may be able to do this. If, for example, teachers attend a session on "Cooperation in Learning Centers," an observer should be able to document the extent to which there is a change in this kind of student behavior over time using a simple observation instrument.

In addition to assessing effects of training on teachers, aides, and students, it is possible to evaluate products resulting from training activities. If part of the in-service program involves materials development, then the resulting materials can be listed, described, and evaluated in terms of their relevance, usefulness, and other features. Sites may choose to evaluate the management of the staff development component in order to provide useful information to improve next year's training program.

The term "staff" can be defined as project staff, or more broadly as all district staff, or even more broadly as staff from other districts. If non-project staff are included in in-service sessions, or if they receive information about the project, then the effects of these efforts can be evaluated and discussed. If the practices employed by the project are so innovative or successful that they are influencing neighboring districts, then this is an important benefit to others resulting from the project.

A number of suggestions are expressed in outline form in the attached framework entitled "Approaches to Evaluation of Staff Development Component." The purpose of the framework is to help explore the variety of

ways in which a district can describe the benefits resulting from staff training.

There are five major sections corresponding to the topics that are underlined above. Within each of these topics suggestions are offered in three areas: (1) the time frame for evaluation, (2) the characteristics assessed, and (3) assessment methods. Each item that is preceded by a bullit (o) is simply a suggestion, and suggestions are not intended to be all inclusive.

The time frame for evaluating staff training can be viewed in several ways. Each event can be evaluated. For example, the content of a workshop, or teacher performance in the classroom can be evaluated after each workshop. Other approaches are to look at changes that occur from fall to spring, from fall to fall, or cumulatively over several years. Some specific suggestions are offered for characteristics to be assessed in the evaluation. These will depend on who or what is being assessed and the nature of the training that was offered. In addition, some assessment techniques are suggested. Measurement is problematic for this program component since it is difficult to obtain valid and reliable measures of changes resulting from training. If it proves unfeasible to employ an assessment instrument of some sort, then simple description should be used. The number of approaches used for evaluating staff development and the extent of their use will depend, of course, on time and financial constraints, but at least program staff can make informed choices based on a number of options.

## Approaches to Evaluation of Staff Development Component

A. Evaluation of content (workshops, presentations, courses, conferences)

1. Time frame
   - for each event
   - over one year
   - over project life

2. Characteristics assessed
   - language of presentations
   - quantity (number of hours per year, etc.)
   - meeting needs of individuals
   - practicality
   - new information
   - expertise of presenter
   - meeting stated objectives
   - relevancy to program needs and resources
   - clarity
   - exchange of ideas
   - continuity
   - variety
   - degree of participant involvement

3. Assessment methods/description techniques
   - rating scale
   - interviews with staff
   - questionnaire for recipients
   - simple description of training

B. Evaluation of effects on instructional staff

1. Time frame
   - each event (ex: one-shot post workshop assessment)
   - over one year
   - over project life

2. Characteristics assessed (depends on nature of training)
   - classroom performance
   - degrees, certification, endorsement
   - knowledge and skills
   - attitudes
   - commitment
   - language skills
   - involvement with parents and community
   - roles of teachers, aides, volunteers
   - self-concept of instructors

- management skills
- evaluation skills
- community

3. Assessment methods/description techniques

- classroom observation
- videotape
- test
- rating scale
- interview
- questionnaire
- tally
- description
- pre-post needs assessment

C. Evaluation of effects on students

1. Time frame

- each event (ex: one shot post-workshop assessment or pre-post workshop assessment)
- over year
- over project life

2. Characteristics assessed (depends on content of training)

- self direction of students
- time on task
- language use
- interethnic interaction
- cooperation in learning centers
- motivation
- student work production

3. Assessment methods/description techniques

- classroom observation
- teacher questionnaire
- teacher interview
- tests
- student interview
- parental report

D. Evaluation of products resulting from training sessions (materials, record-keeping system, etc.)

1. Time frame

- each event
- over one year
- over project life

2. Characteristics assessed

- quantity
- quality
- usefulness
- relevance to curriculum

3. Assessment methods/description techniques

- list and description
- rating scale
- documentation of dissemination
- documentation of extent of use

E. Evaluation of management of staff development component

1. Time frame

- each event
- over year
- over project life

2. Characteristics assessed

- project director's role
- instructional coordinator's role
- adequacy of planning and implementation
- coordination with staff
- cost effectiveness
- inclusion of non-project personnel in project activities

3. Assessment methods/description techniques

- participants questionnaire
- rating scale
- individual interviews

F. Evaluation of effects on non-project staff (including other districts)

1. Time frame

- each event (presentation, mailing, etc.)
- over year
- over project life

2. Characteristics assessed

- knowledge of or awareness of project goals and methods
- degree of coordination between project and non-project classrooms
- attitudes toward bilingual education
- interest in participating in project

3. Assessment methods/description techniques

- questionnaire
- list and description
- record of number of visitors to project
- record of number of requests for information about project from neighboring districts
- documentation of extent of dissemination effort

## EVALUATION OF PARENT/COMMUNITY INVOLVEMENT

### Common Practices

Most Project Information Package (PIP) tryout sites documented and reported events sponsored for or by parents and community members. Whether or not changes came about because of parent/community participation was often not addressed. Little or no attention was given to examining the effects of this component on the school, the students or the community itself. The following evaluation approaches were by far the most common:

- reporting attendance at parent advisory committee (PAC) meetings, and presenting minutes and a list of accomplishments;
- describing parent workshops, parent education sessions, and reporting attendance;
- documenting efforts to disseminate information about the school and the project to parents and community;
- documenting home visits by staff and parent/teacher conferences.

A limited number of sites employed additional evaluation techniques, including the following:

- use of a pre-post questionnaire to measure parents' gains in knowledge of bilingual education, and attitude toward the program;
- documentation of parent activities in the school (as tutors, field trip supervisors, etc.);
- list of products of parent/community workshops (instructional games, cassette recordings, newsletter, etc.);
- parent questionnaire to assess value of their participation in school activities;
- parent questionnaire to assess whether or not information was received about project and about project evaluation;
- questionnaire addressed to PAC to assess strengths and weaknesses of the bilingual education project;
- survey to assess child's home language use.

Recommendations

To a large extent the success or failure of a program is determined by the contextual features which characterize it. Parent/community (P/C) support of a bilingual education program can be a great asset in helping the program gain advocates and support. For this reason the first recommendation is to document and report the type of community support a program received throughout the various stages of program development (planning stage, implementation). The amount of support a program receives initially can be a predictor of the type of support it will receive throughout its life, unless some community feature changes dramatically. Once the schools where the program will be housed are selected, it is recommended that some historical data be collected as to the extent of P/C support that existed prior to the program's inception. This information can be used as a comparison in documenting the change in community support over time.

A second recommendation is that realistic, meaningful short-term and long-term objectives be written which will define the expected school-community relationship. P/C participation in this activity is essential since it will outline their commitment to the school as well as their expectations of the school. Assurances ought to be made that minority P/C participation will occur since this is the target population of the bilingual education project and since compliance with federal guidelines is a goal in itself.

A third recommendation is to plan processes and activities which will produce the desired outcomes specified in the goals. The formation of a PAC, production of an activities calendar, formation of standing committees (for hiring, curriculum, evaluation, etc.) are examples of processes which will achieve some of the short-term goals specified. Parents' actual participation in the classroom and cultural instructional units prepared by parents are examples of processes that may contribute to achieving some of the desired long-term goals.

A fourth recommendation is one that is presently being addressed by most sites. This is to document the array of activities that take place throughout the school year that are of significance to the school-community marriage. Section A-2 of the following outline lists a variety of P/C activities common to bilingual education programs. This list is by no means exhaustive; however, it categorizes activities in a systematic manner so that it is possible to identify the gaps and weaknesses as well as the strengths in a program. P/C activities and characteristics are grouped by domains such as management, curriculum, and parent advisory committee.

A thorough evaluation of the P/C component requires going one step beyond the documentation of activities. It requires an attempt to respond to the questions, "What are the effects of the P/C component on the P/C itself, on the students, and on the school?" The goals mentioned earlier should specify the changes expected to be produced in each of these areas. The next question to ask is, "How will these changes be manifested?" The answer to this question will determine the choice of the assessment method and time frame most appropriate for each area to be evaluated. Sections B, C, and D of the outline address these questions and offer suggestions for selecting characteristics to be assessed, assessment methods, and a time frame.

## APPROACHES TO EVALUATION OF PARENT INVOLVEMENT COMPONENT

### A. Parent/Community Involvement Activities and Characteristics

1. Descriptive Information on P/C Participation

   a. Historical parent involvement at school selected to house the bilingual program (a comparison standard)

   b. Type and amount of community involvement at selection/adoption stage in trying to get the Title VII grant

   c. Amount of time devoted to P/C affairs by staff liaison; source of funds for position

   d. Paid or volunteer positions held by parents (community liaison, teacher aides, etc.)

2. List of Parent/Community Involvement Activities

   a. Management
      - forming staff hiring standing committee
      - planning calendar of school events (holidays, plays, carnivals, open house, etc.)
      - planning student progress reporting procedures

   b. Curriculum planning activities
      - goals and objectives
      - materials selection
      - cultural component
      - first and second language use plan
      - extra curricular activities
      - planning parent classroom participation

   c. Classroom involvement
      - parent function (tutor, clerical, PAC, parent education, etc.)
      - language used in activity (English, other)
      - duration of parent participation
      - contribution of parent (helpful, informative, entertaining, productive)
      - relevancy to program objectives
      - continuity

d. Parent Advisory Committee activities and characteristics

- parent input to PAC constitution and rules and regulations
- officers elected vs. appointed (duration of term, qualifications, appointed or elected by whom)
- responsibilities, powers, and limits of PAC
- participation of project members (numbers, percentages involved in action committees)
- participation of minority as compared with non-minority parents
- participation of community organizations and/or individuals
- parent in-service and parent education
- PAC budget

e. Reporting and evaluation activities

- PAC standing committee on evaluation
- classroom visitation (frequency, duration, purpose)
- parent training on evaluation
- parents' involvement in testing
- PAC's evaluation effort as reflected in yearly evaluation report

B. <u>Evaluation of Effects of Parent Participation on Parents and Community</u>

1. Time Frame

- each day
- each event (curriculum unit, parent educational course, a field trip, etc.)
- pre-post (yearly)
- longitudinal (program's duration)

2. Characteristics Assessed

- participants' performance (as PAC members, tutors, etc.)
- degrees, certification, awards, etc.
- knowledge of project and skills acquired
- attitudes towards project
- commitment (actual participation, support)
- role of parents in school affairs
- self-concept of parents

3. Assessment Methods
   - classroom observation
   - test
   - questionnaire
   - rating scale
   - interview
   - tally

C. Evaluation of Effects of Parent Participation on Students

1. Time Frame
   - each event
   - pre-post year assessment
   - over students' program participation

2. Characteristics Assessed
   - students' change in discipline
   - time on task
   - language usage
   - inter-ethnic interaction
   - motivation
   - student work product
   - attitude
   - absenteeism
   - retentions

3. Assessment Methods
   - classroom observation
   - teacher questionnaire
   - tests
   - student interview
   - parental report
   - count, tally of students' work production

D. Evaluation of Effects of Parent Participation on School

1. Time Frame
   - each event
   - pre-post yearly
   - historical (pre-project current)
   - over project life

2. Characteristics Assessed
   - staff characteristics
   - teachers' classroom performance
   - classroom ambience
   - parent-school communications
   - language usage in school
   - inter-ethnic interaction
   - curriculum appropriateness
   - school budget
   - project evaluation

3. Assessment Methods
   - classroom observation
   - rating scale
   - questionnaire
   - tally
   - description
   - pre-post needs assessment

# REFERENCES

## REFERENCES


American Institutes for Research. Evaluation of the impact of ESEA Title VII Spanish/English bilingual education program, Volume I: Study design and interim findings. Palo Alto, CA: AIR, February, 1977.

Bissell, J. S. Program evaluation as a Title VII management tool. Forthcoming. Los Alamitos, CA: Southwest Regional Laboratory for Educational Research and Development (SWRL).

Bye, T. T. Tests that measure language ability: A descriptive compilation. Berkeley, CA: BABEL/LAU Center, 1977.

Center for the Study of Evaluation. CSE summative evaluation kit. Los Angeles, CA: University of California, 1975.

Center for Applied Linguistics. Bilingual education: Current perpectives (5 vols.). Arlington, VA: CAL, 1977-78.

Dieterich, T. G., Freeman, C., & Crandall, J. A. A linguistic analysis of some English proficiency tests. Paper presented at National Association for Bilingual Education Conference, Seattle, WA, May 1979.

Dissemination and Assessment Center for Bilingual Education. Evaluation instruments for bilingual education: An annotated bibliography. Austin, TX: DACBE, 1976.

Gilmore, G., & Dickerson, A. The relationship between instruments used for identifying children of limited English speaking ability in Texas. Houston, TX: Region I, 1979.

Hoepfner, R. Achievement test selection for program evaluation. In In Wargo, M. J. and Green, O. R. (ed.), Achievement testing of disadvantaged and minority students for educational program evaluation. CTB/McGraw Hill, 1977.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement. Washington, DC: U.S. Government Printing Office, 1975.

Houts, P. L. The myth of measurability. New York: Hart Publishing Company, Inc., 1977.

Hubert, J. An investigation of the Language Assessment Battery (English, Level I) for Title VII students in Hartford. Hartford, MA: 1978.

Law, A. Proceedings of the Bilingual Instrument Review Committee (AB 3470). Sacramento, CA: Office of Program Evaluation and Research, California State Department of Education, September 28, 1978.

Loret, P. G., et al. Anchor test study. Washington: U.S. Government Printing Office, 1974.

Mackey, W. F., & Beebe, V. N. Bilingual schools for a bicultural community: Miami's adaptation to the Cuban refugees. Rowley, MA: 1977.

Northwest Regional Educational Laboratory. Oral language tests for bilingual students: An evaluation of language dominance and proficiency instruments. Portland, OR: NWRL, 1976.

Northwest Regional Educational Laboratory, Center for Bilingual Education. Assessment instruments in bilingual education: A descriptive catalogue of 342 oral and written tests. Los Angeles, CA: National Dissemination and Assessment Center, 1978.

Pletcher, B. P., Locks, N. A., Reynolds, D. F., & Sission, B. G. A guide to assessment instruments for limited English speaking students. New York, NY: Santillana Publishing Company, Inc., 1978.

Rhodes-Hoover, M., Politzer, R. L., & Taylor, O. Bias in achievement and diagnostic reading tests: A linguistically oriented view. Unpublished manuscript, Stanford University, 1975.

Roberts, A. O. H. Thresholds and decisions. Mountain View, CA: RMC Research Corporation, June 1977.

Roberts, A. O. H. Out-of-level testing. Mountain View, CA: RMC Research Corporation, 1978.

Spolsky, B. (Ed.) Advances in language testing series: 1 - Approaches to language testing. Papers in Applied Linguistics. Arlington, VA: Center for Applied Linguistics, 1978.

Spolsky, B. (Ed.) Advances in language testing series: 2 - Some major tests. Papers in Applied Linguistics. Arlington, VA: Center for Applied Linguistics, 1978.

Tallmadge, G. K. Cautions to evaluators. In Wargo, M. J. and Green, O. R. (ed.), Achievement testing of disadvantaged and minority students for educational program evaluation. CTB/McGraw Hill, 1977.

Tallmadge, G. K., & Wood, C. T. User's Guide: ESEA Title I evaluation and reporting system (Rev. ed.). Mountain View, CA: RMC Research Corporation, 1976.

Texas Education Agency. Report from the committee for the evaluation of language assessment instruments, 1977.

Wargo, M. J., & Green, D. R. Achievement testing of disadvantaged and minority students for program evaluation. CTB/McGraw Hill, 1977.